

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Volume X

1950

BOX 6907, COLLEGE STATION, DURHAM, N. C.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

Editor..... G. FREDERIC KUDER

Assistant Editor..... MARCIA M. MATTHEWS

ASSOCIATE EDITORS

DOROTHY C. ADKINS..... University of North Carolina

JOHN H. ROHRER, Editorial Representative of the American College
Personnel Association..... Tulane University

M. W. RICHARDSON..... Richardson, Bellows, Henry and Co.

BOARD OF COOPERATING EDITORS

JOHN G. DARLEY
University of Minnesota

DAVID SESEL
U. S. Office of Education

HAROLD A. EDOERTON
Richardson, Bellows, Henry and Co.

C. L. SHARTLE
Ohio State University

MAX D. ENGELHART
Chicago City Junior College

H. C. TAYLOR
*The W. E. Upjohn Institute for Com-
munity Research*

E. B. GREENE
Wayne University

THELMA G. THURSTONE
University of Chicago

J. P. GUILFORD
University of Southern California

HERBERT A. TOOPS
Ohio State University

E. F. LINDQUIST
State University of Iowa

E. G. WILLIAMSON
University of Minnesota

CHARLES I. MOSIER
Personnel Research Section, A.G.O.

BEN D. WOOD
Columbia University

P. J. RULON
Harvard University

JOHN R. YALE
Science Research Associates

This journal is open to: (1) discussions of problems in the field of the measurement of individual differences, (2) reports of research on the development and use of tests and measurements in education, industry, and government, (3) descriptions of testing programs being used for various purposes, and (4) miscellaneous notes pertinent to the measurement field, such as suggestions of new types of items or improved methods of treating test data. Contributors receive one hundred reprints of their article without charge. Manuscripts should be sent to G. Frederic Kuder, Box 6907, College Station, Durham, North Carolina. Writers are requested to include a biographical sketch with each manuscript, following the style of the section on contributors published in each issue.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT is published quarterly, one volume per calendar year, at Mount Royal and Guilford Avenues, Baltimore 2, Maryland and Durham, North Carolina. Entered as second class matter August 16, 1948, at the Post Office at Baltimore, Maryland, under the Act of March 3, 1879.

Subscription rate, \$5.00 a year, domestic and foreign. Single copies, \$1.50, with the exception of Volume VII, No. 3, Volume VIII, No. 3, and Volume IX, No. 1, for which the price is \$2.50 each. Back volumes: Volumes V (1945), VI (1946), VII (1947), VIII (1948), and IX (1949), \$6.00 each. Volumes I through IV are available in a small-print edition at \$3.00 per volume (paper bound).

Orders should be sent to EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, Box 6907, College Station, Durham, North Carolina.

Copyright, 1950, by G. Frederic Kuder

INDEX FOR VOLUME X

<i>Adams, Sidney</i>	
Does Face Validity Exist?	320
<i>Ammons, Robert H. (with Lee D. Rachler)</i>	
The Full Range Profile Vocabulary Test: II. Selection of Items for Final Scales	327
<i>Anastasi, Anne</i>	
The Concept of Validity in the Interpretation of Test Scores	357
<i>Baker, P. C. (with C. H. Lawther)</i>	
Three Aids in the Evaluation of the Significance of the Difference Between Percentages	383
<i>Balinsky, Benjamin</i>	
Psychological Testing for Immigrants in a Vocational Counseling Agency	378
<i>Beckley, Donald K.</i>	
Problems in Measuring the Effectiveness of Professional Education	37
<i>Berge, William K.</i>	
Preferencia and Behavior Ratings of Dominance	192
<i>Blacker, W. W. (with Clifford P. Froelich)</i>	
Major Issues and Trends in the Graduate Training of College Personnel Workers	388
<i>Berg, Walter K.</i>	
The Interests of Art Students	100
<i>Boynton, Marcia</i>	
Inclusion of "None of These" Makes Selecting Items More Difficult	431
<i>Brody, David S.</i>	
A Genetic Study of Sociality Patterns of College Women	313
<i>Brogden, Hubert E. (with Edwin K. Taylor)</i>	
The Theory and Classification of Criterion Bias	169
<i>Brother Louis</i>	
The Role of Student Government in the Student Personnel Program	169
<i>Burton, Claude E. (with T. W. Holley)</i>	
A Factorial Study of Brethren	400
<i>Callis, Robert</i>	
Change in Teacher Pupil Attitudes Related to Training and Experience	718
<i>Cattell, R. B. (with J. B. Hunt, P. J. Hunt and R. G. Stewart)</i>	
The Objective Measurement of Dynamic Traits	224

<i>Clague, Ewan</i>	EMPLOYMENT OUTLOOK FOR THE 1960s: A GUIDE FOR COLLEGE GRADUATES	167
<i>Cohen, Louis D.</i>	PATTERNS OF RESPONSE IN LEARNING AND ATTITUDE IN LEARNING	173
<i>Coombs, C. H.</i>	THE CONCEPTS OF RELIABILITY AND HOMOGENEITY	181
<i>Cottle, William C.</i>	A NOTE ON THURSTONE'S METHOD OF COMPARING THE INVERSE OF A MATRIX	184
<i>Cronbach, Lee J.</i>	FURTHER EVIDENCE ON RESPONSE SET AND TEST DESIGN	187
<i>Cross, Orrin H.</i>	A STUDY OF FAKING ON THE KUDER PREFERENCE BOARD	195
<i>Cureton, Edward F.</i>	VALIDITY, RELIABILITY AND BIASING	198
<i>Curtis, James W.</i>	ADMINISTRATION OF THE PERSONAL PERSONALITY INVENTORY TO INDIVIDUALS	202
<i>Downie, N. M. (with C. R. Pace and M. F. Foy)</i>	A STUDY OF GENERAL EDUCATION AS A SOURCE OF EMPLOYMENT WITH SPECIAL ATTENTION TO THE COLLEGE GRADUATE	209
<i>Downie, N. M. (with C. R. Pace and M. F. Foy)</i>	THE KNOWLEDGE OF GENERAL EDUCATION AS A SOURCE OF EMPLOYMENT: SYRACUSE UNIVERSITY STUDENTS AS RESPONDENTS	214
<i>Downie, N. M. (with C. R. Pace and M. F. Foy)</i>	COOPERATIVE GENERAL EDUCATION: THE SYRACUSE UNIVERSITY MAGAZINE CURRENT AFFAIRS TEST	218
<i>Downie, N. M. (with C. R. Pace and M. F. Foy)</i>	THE OPINIONS OF SYRACUSE UNIVERSITY STUDENTS ON SOME WIDELY DISCUSSSED CURRENT TOPICS	222
<i>Dressel, Paul L. (with Ross H. Matlen)</i>	THE EFFECT OF CLIENT PERCEPTION ON THE TEST INTERPRETATION	225
<i>Dudek, Frank J. (with Robert W. Klemm)</i>	A FACTORIAL INVESTIGATION OF TEST CONSTRUCTION	231
<i>DuMas, Frank M.</i>	A TABLE AND AN ABAC FOR TEST DATA AND SIGNIFICANCE OF RHO	235
<i>DuMas, Frank M.</i>	EVALUATING PSYCHOMETRIC PROPERTIES	237
<i>Edwards, Allen L.</i>	ON THE USE OF INTERACTIONS AS "FURTHER TESTS" IN THE ANALYSIS OF VARIANCE	243
<i>Ellis, Albert</i>	AN EXPERIMENT IN THE RATING OF IDEAL IDEAL EDUCATION QUESTIONS BY COLLEGE STUDENTS	247
<i>Elton, Charles F.</i>	A STUDY OF CLIENT RESPONSIBILITIES CONCERNING TECHNIQUE OR INTERVIEW OUTCOME	251
<i>Embree, Royal M.</i>	DEVELOPMENTS IN COUNSELING BUREAUS AND CENTERS	256
<i>Fassett, Katherine K.</i>	INTEREST AND PERSONALITY MEASURES OF VETERAN AND NON-VETERAN UNIVERSITY FRESHMAN MEN	260

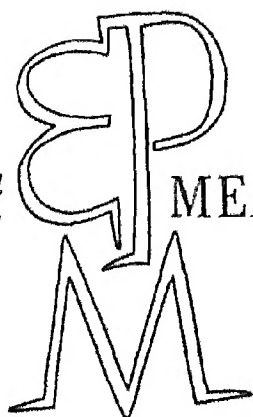
[illegible]

<i>Lauer, A. R. (with William B. Michaels)</i>	
EVALUATION OF AN OPTOMETRIC TEST	182
<i>Lawshe, C. H. (with P. C. Baker)</i>	
THREE AIDS IN THE EVALUATION OF THE SIGNIFICANCE OF THE DIFFERENCE BETWEEN PERCENTAGES	19
<i>Lins, L. J.</i>	
PROBABILITY APPROACH TO FORECASTING UNIVERSITY SUCCESS WITH MEASURED GRADE AS THE CRITERION	196
<i>Mandell, Milton</i>	
MEASURING ORIGINALITY IN THE PHYSICAL SCIENCES	2
<i>Matteson, Ross W. (with Paul L. Dresser)</i>	
THE EFFECT OF CLIENT PARTICIPATION IN TEST INTER- PRETATION	124
<i>McClelland, William A. (with H. Wallace Smarkol)</i>	
AN INVESTIGATION OF A COUNSELOR ATTITUDE QUESTION- NAIRE	225
<i>Michael, William B. (with J. P. Guilford and Wayne S. Lawrence man)</i>	
AN INVESTIGATION OF TWO HYPOTHESES REGARDING THE NATURE OF SPATIAL-RELATIONS AND VISUALIZATION FACTORS	237
<i>Michael, William B. (with A. R. Lauer)</i>	
EVALUATION OF AN OPTOMETRIC TEST	182
<i>Miller, Carroll L.</i>	
DEVELOPMENTS IN COUNSELING BY FACULTY MEMBERS	462
<i>Mills, Thelma</i>	
NO VAIN IMAGININGS	477
<i>Moore, Joseph E.</i>	
THE STANDARDIZATION OF THE MICHIGAN FIVE HAND COLOR DISCRIMINATION AND COLOR MATCHING TEST	379
<i>Myers, R. C. (with D. G. Schultz)</i>	
PREDICTING ACADEMIC ACHIEVEMENT WITH A NEW ATTITUDE- INTEREST QUESTIONNAIRE I	164
<i>Neidt, Charles O. (with Martin F. Fein)</i>	
RELATION OF CYNICISM TO CERTAIN STUDENT CHARACTER- TERISTICS	732
<i>Ohlsen, Merle M.</i>	
DEVELOPMENTS IN RESIDENCE HALL COUNSELING	464
<i>Pace, C. Robert (with N. M. Downie and M. E. Traver)</i>	
A STUDY OF GENERAL EDUCATION AT SYRACUSE UNIVERSITY WITH SPECIAL ATTENTION TO THE OBJECTIVES	169
<i>Pace, C. Robert</i>	
OPINION AND ACTION: A STUDY IN VALIDITY OF ATTITUDE MEASUREMENT	413
<i>Pace, C. Robert (with N. M. Downie and M. E. Traver)</i>	
THE KNOWLEDGE OF GENERAL EDUCATION OF A SAMPLE OF SYRACUSE UNIVERSITY STUDENTS AS REVEALED BY THE COOPERATIVE GENERAL CULTURE TEST AND THE TIME MAGAZINE CURRENT AFFAIRS TEST	264
<i>Pace, C. Robert (with N. M. Downie and M. E. Traver)</i>	
THE OPINIONS OF SYRACUSE UNIVERSITY STUDENTS ON SOME WIDELY DISCUSSED CURRENT ISSUES	423

Portrait of a Man

VALUATION AND STANDARDIZATION OF THE AGO GENERAL MECHANICAL APTITUDE TEST FOR THE SELECTION OF CIVILIAN EMPLOYEES IN WAR DEPARTMENT INSTALLATIONS	254
<i>Ratcliffe, Le. D. (with Robert B. Armon)</i>	
THE FIVE RANGE PICTURE VOCABULARY TEST: II SPECIFICATION OF ITEMS FOR FIFAL SCALES	377
<i>Reynolds, William A.</i>	
NOMOGRAPH OF PETERS AND VAN VOORHIS' APPROXIMATION FORMULA FOR CORRECTING INTERSECTION CORRECTION COEFFICIENTS FOR HETEROGENEITY	137
<i>Rundquist, Edward A.</i>	
RESPONSE SETS: A NOTE ON CONSCIOUSNESS IN TAKING EXTREME POSITIONS	97
<i>Schultz, D. G. (with R. G. Myers)</i>	
PREDICTING ACADEMIC ACHIEVEMENT WITH A NEW ATTITUDE INTEREST QUESTIONNAIRE I	654
<i>Simako, H. Wallace (with William J. McGilland)</i>	
AN INVESTIGATION OF A COUNSELOR ATTITUDE QUESTIONNAIRE	128
<i>Smith, Jr., Robert G.</i>	
REPRODUCIBLE SCALES AND THE ASSUMPTION OF NORMALITY	395
<i>Snider, Harold E.</i>	
OUR STAKE IN THE OCCUPIED COUNTRIES	661
<i>Spache, George</i>	
THE CONSTRUCTION AND VALIDATION OF A WORK-TYPE AUDITORY COMPREHENSION READING TEST	249
<i>Spingaglia, Martin</i>	
AN INVESTIGATION OF THE PERSONALITY TRAITS OF ART STUDENTS	285 ^w
<i>Stewart, R. G. (with R. B. Gault, A. R. Hart and P. A. Hunt)</i>	
THE OBJECTIVE MEASUREMENT OF DYNAMIC TRAITS	224
<i>Strang, Ruth</i>	
MAJOR LIMITATIONS IN CURRENT EVALUATION STUDIES	531
<i>Suchman, Edward A.</i>	
THE LOGIC OF SCALE CONSTRUCTION	79
<i>Taylor, Elwyn K. (with Hubert E. Bragden)</i>	
THE THEORY AND CLASSIFICATION OF CRITERION BIAS	159
<i>Trickett, Robert M. W. (with Winifred L. Wallace)</i>	
THE ASSESSMENT OF THE ACADEMIC ATTITUDE OF THE GRADUATE STUDENT	371
<i>Troyer, M. E. (with N. M. Downie and C. R. Pace)</i>	
A STUDY OF GENERAL EDUCATION AT SYRACUSE UNIVERSITY WITH SPECIAL ATTENTION TO THE OBJECTIVES	359
<i>Troyer, Maurice E.</i>	
PLANS FOR THE NEW INTERNATIONAL CHRISTIAN UNIVERSITY IN JAPAN	603
<i>Troyer, M. E. (with N. M. Downie and C. R. Pace)</i>	
THE KNOWLEDGE OF GENERAL EDUCATION OF A SAMPLE OF SYRACUSE UNIVERSITY STUDENTS AS REVEALED BY THE COOPERATIVE GENERAL CULTURE TEST AND THE TIME MAGAZINE CURRENT AFFAIRS TEST	294

<i>Troyer, M. E. (with N. M. Downie and C. R. Pace)</i>	
THE OPINIONS OF SYRACUSE UNIVERSITY STUDENTS ON SEX	
WIDELY DISCUSSED CURRENT ISSUES	628
<i>Wallace, Wimburn (with Robert M. H. Travers)</i>	
THE ASSESSMENT OF THE ACADEMIC ABILITY OF THE GRADUATE STUDENT. . .	671
<i>Wrenn, C. Gilbert</i>	
AWARD IN STUDENT PERSONNEL RESEARCH	142
<i>Wrenn, C. Gilbert (with Robert B. Kamm)</i>	
CLIENT ACCEPTANCE OF SELF-INFORMATION IN COUNSELING.	12
<i>Zimmerman, Wayne S. (with William B. Michael and J. P. Gaultford)</i>	
AN INVESTIGATION OF TWO HYPOTHESES REGARDING THE NATURE OF THE SPATIAL-RELATIONS AND VISUALIZATION FACTORS.....	107



VOLUME TEN, NUMBER ONE, SPRING, 1950

<i>Further Evidence on Response Sets and Test Design.</i> LEE J. CRONBACH.	3
<i>Client Acceptance of Self-Information in Counseling.</i> ROBERT B. KAMM AND C. GILBERT WRENN.....	32
<i>The Concepts of Reliability and Homogeneity.</i> C. H. COOMBS..	43
<i>Problems in Measuring the Effectiveness of Professional Education.</i> DONALD K. BECKLEY.....	57
<i>The Concept of Validity in the Interpretation of Test Scores.</i> ANNE ANASTASI.....	67
<i>The Logic of Scale Construction.</i> EDWARD A. SUCHMAN.....	79
<i>Validity, Reliability and Baloney.</i> EDWARD F. CURETON.....	94
<i>Response Sets: A Note on Consistency in Taking Extreme Positions.</i> EDWARD A. RUNDQUIST.....	97
<i>The Interests of Art Students.</i> WALTER R. BORG.....	100
<i>A Factorial Investigation of Flexibility.</i> ROBERT W. KLEEMEIER AND FRANK J. DUDEK.....	107
<i>The Standardization of the Moore Eye-Hand Coordination and Color Matching Test.</i> JOSEPH E. MOORE.....	119
<i>An Investigation of a Counselor Attitude Questionnaire.</i> WILLIAM A. MCCLELLAND AND H. WALLACE SINAIKO.....	128
<i>A Note on Thurstone's Method of Computing the Inverse of a Matrix.</i> WILLIAM C. COTTLE.....	134
<i>Nomograph of Peters and Van Voorhis' Approximation Formula for Correcting Interfunction Correlation Coefficients for Heterogeneity.</i> WILLIAM A. REYNOLDS.....	137
<i>A Single Chart for Tetrachoric r.</i> WILLIAM LEROY JENKINS...	142
<i>New Tests.</i>	145

FURTHER EVIDENCE ON RESPONSE SETS AND TEST DESIGN

LEE J. CRONBACH¹

University of Illinois

WHEN a person takes an objective test, he may bring to the test a number of test-taking habits which affect his score. Personal ways of responding to test items of a given form (e.g., the tendency to say "agree" when given the alternatives "agree" "uncertain" "disagree") are frequently a source of invalidity. In 1946, the writer (4) assembled evidence demonstrating that these "response sets" are present in a wide variety of tests. Since that time, much new evidence has come to light, and it is now possible to examine more completely the nature of response sets. While much of the material to be reported is new, evidence has also been drawn from scattered publications which were overlooked in the earlier review. Material on response sets is to be found in a great many sorts of studies, discussed under many names. Particular attention should be drawn to the early reports of Lorge (15) and Goodfellow (6) on this topic.

As our earlier report demonstrated, response sets have been identified in tests of ability, personality, attitude, and interest, and in rating scales. Among the most widely found sets are acquiescence (tendency to say "True," "Yes," "Agree," etc.), evasiveness (tendency to say "?," "Indifferent," "Uncertain," etc.), and similar biases in favor of a particular response when certain fixed alternatives are offered. Other sets include the tendency to work for speed rather than accuracy, the tendency to guess when uncertain, the tendency to check many items in a checklist, etc. Response sets become most influential as items become difficult or ambiguous. Individual differences in response sets are consistent throughout a given test, as shown

¹This study was assisted by funds from the Bureau of Research and Service, College of Education.

by split-half coefficients. Response sets dilute a test with factors not intended to form part of the test content, and so reduce its logical validity. These sets may also reduce the test's empirical validity. Response sets tend to reduce the range of individual differences in score.

The pattern of this discussion is as follows: First, many studies are cited which bolster the conclusion that response sets are widely found, and are particularly influential when a test is difficult. These new sources confirm earlier findings and do not modify them. The significant new material in this section relates to two multiple-choice tests, and confirms the hypothesis that this form of test is nearly free from response sets. The second section of the report deals with the nature of response sets. Questions considered are: Can performance be altered by special directions or training to avoid response biases? Are response sets consistent traits, so that a person shows a similar set on different tests? Are response sets correlated with other aspects of personality? These studies deal particularly with the question whether response sets are due to a transient mind-set and are therefore only a nuisance in testing, or whether they may provide data on important variables. The third and final section reviews methods used to control the influence of response sets on validity, and discusses what test constructors can do to design better tests.

Evidence that Response Sets Exist

It is scarcely necessary to marshal further evidence that reliable individual differences in response sets exist. Yet the widespread use of test forms which permit responsesets indicates that their existence is not adequately appreciated. It is not only the old tests—Seashore, Bernreuter, Thurstone attitude, Strong—that suffer from response sets. New tests appear continually, especially tests of attitude and personality, whose forms invite response sets. The writer has routinely requested graduate students to analyze their data for response sets whenever their research employed tests with fixed response categories (A-U-D, Yes-No-?, etc.). *Never has such an analysis failed to disclose individual patterns of response, statistically consistent from item to item.*

The most effective simple design to demonstrate response sets is to obtain a score for each person on the suspected response set. Thus, Lorge tested the existence of "gen-like," or acquiescence on the Strong test, by counting how many items each person marked "I." The split-half or Kuder-Richardson reliability of the response-set score can then be computed. Table 1 condenses the evidence obtained by this and other techniques, evidence which, together with that previously assembled, shows conclusively that response sets are to be found in a great many tests.

One study requires a separate report, because it is based on a factorially-designed test in which items are intended to be homogeneous. Kenneth Fells supplied the writer with tests "Cards" and "Figures," from Thurstone's *Tests of Primary Mental Abilities*, which had been given to pupils in a Mid-western city as part of a study by the University of Chicago Committee on Cultural Factors in Intelligence Tests, under a grant from the General Education Board. Both of these tests present a geometric figure at the left of the row, and follow it with figures just like the given one save that they have been rotated through 90° , 180° , or 270° , or are mirror-images of one of these rotations. Directions are to "mark every card (figure) that is like the first card (figure)." It was observed that some pupils seem to search for all correct answers, whereas others are content to identify one or two seemingly correct answers, and then go on to the following row. Papers were drawn at random from those given to all pupils in two large junior high schools. Papers were discarded where any row had been omitted, or where the total score on Cards was high (46 or more out of 54 possible). This avoids spuriously high apparent reliability for the response-set score. The test had been given with double time, and the test was in effect unspecced for the pupils studied. Two response-set scores were obtained for each pupil: Cards $R + W$, and Figures $R + W$. This score indicates a tendency to mark many items in a row. It implies thoroughness and persistence in marking, and perhaps acquiescence. The correlation of the two $R + W$ scores is .54 ($N = 109$). On the whole, those who mark fewer items appear to be poorer students, but no estimate of response sets, independent of ability, could be obtained. For the selected cases, the correlation of $R + W$ cards with $R - W$ Figures was .44, and that of $R + W$ Figures with $R - W$ Cards was .33. These data are interpreted as showing that in addition to the space factor (ability to discriminate similar forms), performance on this test is influenced by a response set. Many students are found who mark few or no incorrect figures ($R + W = R - W$) but who fail to mark all

TABLE I
Studies Reporting Response Sets

Investigator and reference	Name and nature of test	Response called for	Response set	Finding
Bennett, Seashore, Wesman (1)	Differential Aptitudes, Clerical	Checking errors	Speed vs carefulness	Good students may earn falsely low scores due to set to work accurately at slow speed.
Brotherton, Read, Pratt (2)	Questionnaire on word meanings	Checking fixed categories on six-point scale	Definition of terms	Substantial differences in meaning are found from person to person and group to group. Questionnaires involving many, few, several, etc., "are invalid and unreliable."
AAF (7)	Tests of plotting, scale reading, etc.	Solving many items, with time limit	Speed vs carefulness	In one test, reliability of Rights .76, Wrongs .56. But intercorrelation only -.48. Factor analysis shows "carefulness" often the most prominent factor in Wrongs scores.
Humm and associates (11)	Humm-Wadsworth temperament	Yes-No	Acquiescence	High No-Count reduces validity of scores.
Knoell*	Spelling	Check words correctly spelled, and respell those given incorrectly; also spell from dictation	Acquiescence, tendency to omit	Factorization of twenty scores, for tests of same and different types, applied to sixth-graders in Indian schools, shows four factors having low intercorrelations. These are: (1) general ability to spell, (2) ability to recognize correct spellings, (3) tendency to mark many items "right," (4) tendency to omit dictated words. Av. loadings in the respective factors for correct-spellings-marked-correct are .06, .46, .93, -.04. Loadings for incorrect-words-checked-correct are -.82, .25, .74, .02. For difference of these scores, .86, .13, -.04, -.10. In twelfth-graders, acquiescence factor does not appear but recognition of correct spellings remains distinct from general ability.
Longe (15)	Strong interest	L-I-D	Acquiescence, evasiveness	Reliability for number of L's in two testings is .8, for number of I's, .84.

Longe (15)	Thurstone attitude	Checking "agrees"	Acquiescence, evasiveness	Reliability for number of checks in two testings is 88; for number of P's, .95
Mathews (16)	Interests	L-I-I-d-D	Acquiescence	Reliability of tendency to "like" many items is .75-.79 Responses are altered when choices are in order D-d-I-I-L. Responses at extreme left and fourth from left tend to be used. Shift is greatest on items where students have least pronounced views.
Philip (19)	Judgment of proportion in color mixtures	Absolute judgment on 11-point scale	Tendency to use certain portions of scale	Some individuals scatter their judgments more broadly over the scale than others. Each individual uses certain "foci" along the scale more often than other responses Stimuli at the foci and ends of the scale are more often judged correctly than others "Subsidiary cues" have greatest influence when discrimination is difficult.
Rubin (20)	Seashore pitch	H-L	Tendency to judge H	Score on High items only has K-R reliability .827, on Low only, .790 Correlation High x Low only .265, so two types of item do not measure same factor Bias increases with difficulty Bias score (H-L) has reliability .725 †
Singer and Young (21)	Judgments of pleasantness of stimuli	Rating on continuous scale	Tendency to rate P	"When definite affective reactions not aroused, subjects show habitual ways of using rating scale" Differences stable over two weeks.
Thorndike (22)	Pressey Interest-Attitude	Checking worries, interests	Checking many items	Frequent individual differences, reliable by split-half method Low checking threshold leads to low emotional maturity score
Vernon (24)	Interest items	Checking likes	Checking many items	Twenty per cent of variance appears in this general factor, running through all items regardless of content
Wesman (26)	Spelling	Check all misspelled words	Acquiescence	Incorrect spellings correlate higher with total test than correctly-spelled items

* Part of incomplete study, to be published later

† Computed by the writer with Rubin's help

the correct alternatives. Since some of the reliable response-set variance is uncorrelated with the space factor, the entrance of response sets reduces the factorial purity of the test. Certainly tests which aim at measurement of a single factor must be designed to eliminate response sets.

Response Sets in Multiple-Choice Tests—The only major form of fixed-alternative test which has so far been found free from response sets is the multiple-choice item. In order to determine whether response sets can be extracted from a typical test of this type, the writer has studied the *Henmon-Nelson Test of Mental Ability, Form A*, for Grades 3-8. The data for this study were supplied by Eells, from the study which provided the Thurstone data discussed above. Thousands of test papers were available, since every child in several grades in a mid-western city had been tested. The sample for this study was chosen indiscriminately, from papers of upper-lower and lower-middle-class children. In administering the test experimentally, Eells allowed an extended time of 20 minutes beyond the standard time of 30 minutes. Papers not completed even in the extended time were discarded in the present analysis.

The Henmon-Nelson is a suitable test for investigating response sets because items were prepared with care, are fairly well arranged as to difficulty, and are designed so that the correct answer appears about equally often in each of the five response-positions. The hypothesis is that some students may persistently tend to select choices early in the group of five. This would raise their scores on items where the correct answer is choice "1" or "2," but lower than on items keyed "4" or "5." The psychological basis for the hypothesis is the possibility that some students read every alternative and discriminate carefully, where some merely read through the item to find a plausible answer, mark it, and go on to the next item.

The procedure was the usual one: to obtain a "bias" score for each individual and determine its reliability. If the score is reliable, the response set is proved to exist. The response set score for the present hypothesis consists of "number of errors appearing to the left of the correct answer" minus "number of errors to the right of the correct answer." Before rescoring papers for bias, papers of high-scoring pupils (those having

a score above 60 out of 90 items correct) were discarded. This was done to *increase* the likelihood of finding a response set, since response sets have no opportunity to show themselves when the pupil gets most items correct. For a group of 66 papers, bias scores ranged from 24 to -12. The person with the bias score 24 had made 39 errors to the left of the true answer, and only 15 errors to the right of the true answer. Such a preponderance is hard to explain as other than a habit of marking items. For the cases studied, however, the split-half reliability of the bias score was only .095, corrected. Such a low correlation indicates that the postulated response set is of no consequence for this group. A second sample of 84 cases having raw scores of 40 or below in extended time (these pupils had IQ's near or below 80) were studied separately, in order to increase the probability of finding a response set. For these pupils, the reliability of the bias score was .42, corrected. Evidently for a group of pupils taking a difficult multiple-choice test, reliable response sets can be found. Bias has a slight relation to raw score; the mean raw score for these poor pupils was 24.5 for those with negative bias, and 29 for those with positive bias. For some reason, very poor students tended to mark alternatives to the right of the correct answer proportionately more often than slightly better pupils.

An attempt was made to demonstrate such biases as "preference for position 1." No statistical evidence for such sets could be obtained, although an occasional case does suggest that such biases may occur. One boy, for example, never in 90 items marks the fifth choice as correct, and another student places 30 of his marks on position "1."

A second study was made with a modified version of the *Ohio State University Psychological Examination*, using data made available by N. L. Gage and Dora Damrin. The shortened test they used consists of 90 five-choice vocabulary items, unspecced. This test was administered to unselected juniors and seniors in several high schools. When papers for all 171 pupils were scored for tendency to place answers before rather than after the correct position, the odd-even reliability of the bias score was found to be .20. When only the lowest 65 students (as judged by the total number right on the test) were used

as a sample to determine the reliability of the bias score, the reliability rose to .29. This was a group of students for whom the test was extremely difficult; the highest score for the group was 22 right out of 90. It should be noted that this test is normally used for predicting college success among superior high school students; the highest score in this limited subdivision of our sample is only chance expectation. When an even more restricted sample was used—the lowest 26 cases, all of whom fell below a raw score of 15 items correct—the reliability of the bias score rose to .54. The mean bias score changed as the quality of students became poorer. For the total group, the mean bias score was -6.5 ; for the second group, -7.7 ; and for the very lowest group, -9.7 . Here, also, the poorest students apparently tended particularly often to mark errors to the right of the correct answer.

Both of these studies demonstrate that response sets are a minor factor, since so great a selection of cases was required in order to demonstrate any evidence of bias. Probably other multiple-choice tests where all subjects mark all items suffer little from response sets. Confirming studies on other multiple-choice tests are desirable, but the generally satisfactory experience with forced-choice tests should encourage their continued widespread use.

Stability of Response Sets

While there is ample evidence that response sets are consistent throughout a single test, it is important to determine whether they are characteristics of the individual stable from time to time, or are transient sets which can only be regarded as errors in testing rather than personality characteristics.

Some evidence that response sets are stable appears in scattered studies. Thorndike (22, p. 33) reports that on a speeded Air Force test, scores obtained at the same sitting correlate no more than scores obtained several hours apart. If a speed-accuracy set is operating, it is not a set which shifts from hour to hour. Singer and Young (21) found that a tendency to rate varied stimuli as "pleasant" was highly stable, correlations as high as .90 being found under certain conditions over time intervals of two weeks.

Whereas these and similar studies tend to stress the stability in response sets, we ordinarily think of mental sets as easily changed by suitable directions. If the response set is viewed as a way of interpreting an ambiguous situation, as when the word "like" is left for the subject to define, any change in directions should re-define the stimulus elements and alter individual response sets. Several studies show that this can be done.

Rubin (20) several years ago demonstrated the existence of bias in the *Seashore Pitch Test*. He gave the Revised Test B twice to 245 college students, and found that the group as a whole used 13958 "H" responses and only 10542 "L" responses, in judging whether the second tone was higher or lower. According to the key, there were actually an equal number of differences in each direction. A similar mean bias was found by Rubin in data of Farnsworth.

In two ingenious studies Rubin then established that temporary sets are a major element in bias. First he gave a "guessing" test, in which subjects imagined a tossed coin, and wrote down the way they imagined it would fall. One group was given directions as follows: "Imagine a coin which has an *H* for *High* on one side, and an *L* for *Low* on the other side." In the other group this was reversed: "Imagine a coin which has an *L* for *Low* on one side, and an *H* for *High* on the other side." There was a significant preponderance of the first-mentioned response on the first guessed item (i.e., the former group tended to say "H"; the second group to say "L"). There was a significant preponderance of the second-named response on the third guess of the series. Rubin then applied the same reversal to the *Seashore* test directions. 272 students were told, "If the second tone is lower than the first tone, print *L*; if higher, print *H*." Only 56.8 per cent of the errors were lows marked "*H*," compared to 60.0 per cent when much the same group were given the original directions (but note that some bias remained).

A miniature experiment performed by graduate students as a class exercise gives further indication that response sets are easily altered. Lynn Henderson and Esther Williams administered the revised *Seashore Pitch Record B* to ten students, repeating the Record to make a total of 100 items. At the next class meeting, each student's scored paper was returned to him

for brief study. His attention was drawn specifically to the nature of bias by having him count whether he tended to mark "H" more often than "L." He was informed that in each group of ten items, just half were correctly answered "High." The writer conducted the discussion, talking about bias for about fifteen minutes and suggesting strongly that bias could be eliminated with effort and that pitch scores would be improved as a result. Papers were collected as soon as bias had been examined, to reduce the possibility of learning specific answers. Students were never informed, and few suspected, that the same record was used for both items 1 to 50, and 51 to 100. After the discussion,

TABLE 2
Results of Pitch Tests before and after Discussion of Bias

Student	Score on successive tests				Bias on successive tests*				Total score		Total bias	
	IA	IB	IIA	IIB	IA	IB	IIA	IIB	I	II	I	II
1	41	44	42	46	-2	-8	-4	10	84	64	-10	-4
2	45	39	41	43	-2	10	2	-2	84	84	0	0
3	40	42	44	45	-8	-4	8	-2	82	80	-12	6
4	35	41	35	33	2	-6	2	2	76	68	-4	4
5	38	36	29	32	4	-4	2	4	74	64	0	2
6	31	37	31	48	-14	2	2	12	68	70	-12	2
7	30	32	42	39	0	0	-4	-2	62	81	0	-6
8	24	31	32	36	4	2	0	0	48	68	6	0
9	24	27	33	32	-24	-26	10	-8	52	48	-10	-18
10	26	21	30	28	-12	-6	0	0	47	48	-18	0
Median	33	36½	34	37½	-2	-4	0	0	71	73½	-7	0
Mdn. absolute value					4	5	2	2			7	3
Mean	33	35	36	38					68	74		

* Bias score equals number of items marked High minus number marked Low.

the 50-item record was readministered, the papers collected, and the record readministered again, yielding a 100-item post-test. This is admittedly an inadequate experiment, especially in the absence of a control group to measure the effect of practice and suggestion, separated from training regarding bias. The results are nevertheless striking (Table 2). Bias was notable on Tests IA and IB, largely eliminated on IIA and IIB. Total scores generally rose, especially on IIB. The amount of gain in score corresponds somewhat to the amount of initial bias, except for case 7, whose gain is presumably an effect of practice or motivation. This finding is not statistically significant.

This study, small as it is, seems to show that bias can be

eliminated by direct coaching which makes the subject aware of his own bias. If the Pitch Test measured pitch threshold alone, increased insight into habits of responding would not affect scores. The study does not prove that training in bias raises pitch scores, but it strongly suggests that this is true. Wyatt (28) also reports training subjects to avoid bias as a means of improving discrimination. Surely, on the basis of these data, it can be recommended that Seashore test papers should be checked for bias, and that where the person shows a marked bias in either direction scores should be regarded as probably giving too low an estimate of the person's ability to discriminate pitch.

Another report that altering directions affects response sets is made by Goodfellow (6). He finds that in psychophysical judgments the predisposition to report a stimulus as absent was reversed when the directions were worded: "Remember that in approximately one-half of the trials the correct answer will be yes."

The resemblance between response sets inferred from statistical data and "learning sets" found experimentally by Harlow (9) should be pointed out. In studies of monkeys, and also of children, he established definite evidence of generalized learning to solve problems. The monkey enters an ambiguous situation, namely, a discrimination apparatus where the proper choice among two alternatives leads to a food reward. In this situation, a personal communication from Harlow informs us, the monkey demonstrates a preference for one or another of the choices offered (e.g., for the red object rather than the blue). This preference may serve to increase errors (if, for instance, the square object has been keyed as correct, regardless of color). If the monkey is put through one learning series after another, in which a different cue differentiates the right and wrong choices in each series, the monkey quickly learns to learn. His learning curve on later series is strikingly steep. "With each successive block of problems the frequencies of errors attributable to these factors [one of which is initial preference or response set] are progressively decreased. . . . The process might be conceived of as a learning of response tendencies that counteract the error-producing factors."

Harlow has therefore shown that response sets are present in the new, ambiguous situation, and that under his conditions they are extinguished. In contrast, the test-taking sets of adults appear not to be extinguished by usual experiences, even though they increase the probability of error. The difference appears to be that in Harlow's experiment there is an immediate frustration attached directly to the wrong (preference determined) response. In school tests the penalty is delayed, and is usually attached to the total test performance rather than to the specifically wrong responses. False approaches to problems, such as biases, can be eliminated; sound sets, such as reading

TABLE 3
Correlation of Response Sets on Varied Tests

Investigator	Tests	Response Set	Findings
Lorge (15)	Bernreuter, Thurstone attitude, Strong	Acquiescence	Average intercorrelation of number of Yes's .24.
		Evasiveness	Average intercorrelation of number of I's or P's .43
Singer-Young (21)	Two series of tones	Tendency to rate "pleasant"	r's .56, .67.
	Two series of words	Tendency to rate "pleasant"	r's range .44 to .59.
	Two series of different stimulus-types	Tendency to rate "pleasant"	r's range -- .24 to .36
AAF (7)	Wrong's score on four tests of plotting, etc.	Carefulness vs. speed	r's range .14 to .41

each item carefully, can be learned. But direct and immediate teaching will be more effective than such incidental punishments as low total scores.

Generality of Response Sets

To some degree, a person shows consistent response sets from situation to situation. Table 3 summarizes studies bearing on this question. When similar situations are presented, response set scores are significantly correlated. But there is no evidence that response sets are consistent over widely different situations, and Singer and Young's evidence indicates that this is not true. But one does not measure response sets alone. Response sets show only when the response to a situation is in

some way unclear. Singer and Young point out that habits of using their rating scale are operative only when "affective arousal is weak or absent." Perhaps affective arousal is weak for one person on tones, for another on odors. This would reduce the response-set correlations.

Response sets might be mere incidental sources of error in measurement, or they might reflect deeper personality traits. Evidence from many sources now combines to show that response sets reflect "real" variables.

Johnston (13) gave the *Bernreuter Inventory* and the *Hunter Attitude Scale* to two groups of teachers. These groups were chosen on the basis of ratings by their principals, so that one group consisted of "autocratic" teachers, and one consisted of teachers who were markedly "democratic" in classroom practice. Johnston found that these groups differed significantly in response sets. On the Bernreuter, the autocratic group gave an average of 52.6 "Yes," 62.3 "No," and 10.8 "?" responses. The three totals for the democratic group were 55.9, 66.8, and 4.7 respectively. There were 42 teachers in the former group, and 43 in the latter. The difference in "tendency to use question marks" (evasion?) was significant ($P < .01$). There was a similar difference on the Hunter scale. The mean number of statements marked "Undecided" rather than "Agree" or "Disagree" was 15 in the autocratic group and 10 in the democratic group ($P < .01$).

Mersman (17), in a small study of vocational interests, compared the Bernreuter responses of college students planning to be lawyers, musicians, and engineers. There were seventy-five cases in each group. Upon analyzing the number of responses of each type in each group, he found the following means:

	Yes	No	?
Lawyers.....	53	62	10
Musicians.....	56	58	11
Engineers.....	54	64	7

The differences between engineers and musicians are significant (1% level).

Evidently groups differentiated on external criteria also differ in response sets. Where this is so, part of the response-set variance must represent some real variable. For example, use

of question marks may indicate anxiety and evasiveness of personality, rather than a transient set alone. Lorge (15) finds that the tendency to say "Yes," "No," and "?" (estimated from several tests) correlates as follows with scores on the Flanagan-Bernreuter keys:

	Yes	No	?
Confidence....	.27	-.15	-.03
Sociability.....	.00	.27	-.26

Possible significance of response sets for empirical prediction is suggested by a study which finds that tendency to respond "?" is correlated negatively with success in selling life insurance (14). While the relationship found was not statistically significant, the difference between the mean number of question marks in the good and poor groups (8.4 vs. 12.8, CR 1.57) is large enough to suggest further investigation along this line.

Improvement of Test Design

The heterogeneous bits of evidence pieced together here and in our previous report have established several generalizations.

1. Any objective test form in which the subject marks fixed response alternatives ("Yes"- "No," "True-False," "a"- "b"- "c," etc.) permits the operation of individual differences in response sets. The influence of response sets in the multiple-choice test is, however, of minor importance.

2. Response sets have the greatest variance in tests which are difficult for the subjects tested, or where the subject is uncertain how to respond.

3. Items having the same ostensible content actually measure more than one trait, if response sets operate in the test. This is true even for tests which, scored as a whole, are "factorially pure."

4. Slight alterations in directions, or training in test-taking, alter markedly the influence of response sets. But if the situation is not re-structured by the tester, individual differences in response set remain somewhat stable when similar tests are given at different times.

5. Response sets are to a small degree correlated with external variables such as attitudes, interests, and personality. This shows that they are in part a reflection of "real" and

stable traits. To this degree, response-set variance may be valid variance in some investigations.

6. Tests are usually constructed to measure a trait defined by the content of the test items. If the form of the items permits response sets, two persons having equal true scores on the content factor will often receive different scores on the test. Response sets therefore ordinarily dilute the test and lower its validity.

Paragraphs (5) and (6) crystallize the paradox response sets present. Some of the response-set variance is potentially useful, some of it is an interference with measurement. The problem for the tester is to capitalize on the effect of response sets where they are helpful to validity, and to eliminate their influence where it is undesirable. It is therefore important to decide which view is to be taken in any given situation. The writer has attempted to formulate rationally the response-set problem in factorial terms. The analysis has been unsuccessful, primarily because response sets do not obey the fundamental additive law of factor theory. One cannot define a person's test score as a weighted addition of his content-factor and response-set-factor scores, since response sets have an influence on his performance on each item proportional to his doubtfulness. That is, the weight for the response-set factor in any item is not a constant for all persons, but is a function of each person's score in the content factor. Since the problem is not at present formulated analytically in a way which clarifies our thinking, we are confined to a general description of the relations.

Considering only biases such as acquiescence and evasiveness, response-set variance may be conceived as containing the following elements, combined in some proportion:

1. Chance variance; resulting from purely random excess of choice of one or another alternative.
2. Internally consistent but momentary response tendencies; sets operating throughout one testing, but shifting on a retest at another time.
3. Stable response tendencies; sets operating consistently even when the same test is given at different times.

Evidence of the existence of Type 3 variance has been consistently found whenever investigators have sought it. Evidence

for Type 2 variance is lacking, but it may be postulated on the grounds that no observed trait is expected to be perfectly stable. And of course chance variance is always with us.

Response-set variance of Type 1 is not important; it is simply another manifestation of error variance, and its influence can be reduced by lengthening the test. Variance of Type 2 is unquestionably harmful, unless one happens to be doing research on evanescent sets or moods or some other fluctuating variable (for example, a study of mood changes concomitant with fatigue). Type 2 variance cannot correlate with stable variables, and therefore lowers the validity coefficient of the test. Moreover, Type 2 variance is present in many items and probably increases the coefficient of equivalence (split-half or Kuder-Richardson reliability) of the test. Therefore, even if the test given on a particular day were lengthened indefinitely, we could not raise its empirical validity to 1.00 because scores are partly saturated with an invalid factor. Type 3 variance is potentially useful, but to understand its action we must divide it between

- 3a. Valid variance, the portion of 3 that correlates with the criterion the test is intended to predict, and
- 3b. Invalid variance, the portion of 3 that does not correlate with the criterion.

We may always expect a portion of Type 3b, since the response set could correlate perfectly with the criterion only if the criterion is itself a set or a personality trait causing the set.

Variance of Type 3a does exist, since in some studies the response-set score did correlate with some external variable. Moreover, research in a good many fields is turning to personality variables which may be close cousins to response sets. Guilford anticipates that the "carefulness" factor, which is a response-set, may prove to have validity as a component of a battery for aircrew selection. In studies of prejudice or liberalism, an investigator may find evidence on negativism useful. And this is possibly one source of bias toward "No" and "Disagree" in taking tests. Variance of Type 3b reduces validity, and limits the maximum possible validity the test can have even if trials on different days are combined. Variance of Type

3a may increase validity if it is added into the score in one way, or it may lower validity if it is added in differently. Thus the studies of true-false tests (5) show that students tend to say "True" when in doubt, and the duller students, who are in doubt most often, say "True" most often. This raises their score on true items, lowers it on false items. Hence the potentially valid portion of the response-set variance lowers the discriminating power and validity of true items, and enhances the validity of the false items.

Finally, it should be noted that there is no possibility of separating the four types of response-set variance in data from a single test; they come entangled in a single performance, and we must therefore consider the effect of the response-set variance as a unit. This total is made up of a random element (Type 1), a real but invalid element (Type 2, 3b), and a potentially valid element (3a) which may in practice raise or lower the validity of the test score. Of these three categories, only 3a, the valid variance, is likely to be entirely absent, and the size of the correlations of response sets with external variables suggests that 3a is not likely to be the principal component of the variance. Therefore:

- a. The probable effect of response-set variance is harmful, since elements 2 and 3b are usually present, and these elements reduce the extent to which the test is saturated with the content factor it is supposed to measure.
- b. Even if valid variance is present, its effect may be to lower validity of some items or of the total score. But under certain circumstances, it may be treated in such a way that it raises the validity coefficient.
- c. Only under exceptional circumstances, when a test is designed to study the very personality characteristics which are reflected in the response set, does the response set appear to be a potentially helpful source of variance.

Because the operation of response sets upon score is complex, a detailed illustration seems worthwhile. A spelling test is planned, using the directions: "Some of these words are correctly spelled and some incorrect. Mark every item, + if correct, o if incorrect." If the test is intended to indicate whether the student will identify errors in his own writing outside of school, this form of item has an appealing resemblance to the

criterion task. Now suppose we have 6 students. A, B, and C know 40 words out of 60, are doubtful on the remainder. D, E, and F know 30 words. (This oversimplification of "knowing" a word avoids difficulty in this explanation. A) and D have no response set. Of the 60 words, just half are wrongly spelled, and when A and D are doubtful, they mark just half of the unknown words 0. B and E are a little undercritical in non-school writing; they fail to notice some errors. But in taking a school test, they suspect the teacher of planting errors where there are none, and so mark 0 60 per cent of the time when they are doubtful. C and F are undercritical in all their writing, and in taking the test they are also willing to accept errors; they mark 0 only 30 per cent of the time when in doubt. The scores then may develop as follows:

	A	B	C	D	E	F
Bias (Proportion of + responses to 0 responses)....	50/50	40/60	70/30	50/50	40/60	70/30
Words known.....	40	40	40	30	30	30
Guesses correct by chance:						
guessed +.....	5	4	7	7½	6	10½
guessed 0.....	5	6	3	7½	9	4½
Most probable score.	50	50	50	45	45	45
Maximum possible correct guesses:						
guessed +.....	10	8	10	15	12	15
guessed 0.....	10	10	6	15	15	9
Maximum possible score... ..	60	58	56	60	57	54
Minimum possible correct guesses:						
guessed +.....	0	0	4	0	0	6
guessed 0.....	0	2	0	0	3	0
Minimum possible score.....	40	42	44	30	33	36

In this, as in other problems, the tendency is for bias to restrict the range of scores, not to alter the mean score. Where an unbiased person may, with lucky guesses, earn a very high score, the biased person has a much smaller probability of reaching the same total. Bias which reflects "true criticalness" operates in the score no differently from bias which is only a special set used in taking a test. If the items are divided so that 70 per

cent of the words are correctly spelled, C and F are given an advantage, even over A and D. If more than half the spellings are incorrect, B and F will tend to earn higher scores than those who know an equal number of words (and are equal on the criterion).

In an unbiased test, where all alternatives have an equal weight in the total test, response sets do not add to the variance of scores, but have a damping effect, reducing the range of points people may earn from a combination of guessing and partial knowledge. If one alternative is present more than another, response sets form part of the variance of the test scores.

Methods of eliminating response-set variance.—The writer concludes that as a general principle, the tester should consider response sets an enemy to validity. Even when seeking to measure a trait resembling a response set, one can have confidence in the meaningfulness of the score only after showing that variances 1, 2, and 3b are small in proportion to 3a. Therefore, in most tests and certainly in those not intended to measure personality, we should keep response sets from affecting the test score by one of the following methods: designing test items which prevent response sets, altering directions to reduce response sets, or correcting for response sets.

(a) *Test design.*—Since response sets are a nuisance, test designers should avoid forms of items which response sets infest. This means that any form of measurement where the subject is allowed to define the situation for himself in any way is to be avoided. (We must make an exception for tests where his way of interpreting the test is treated as a significant variable. But even so, the above analysis suggests limits to the possible validity of tests like the Rorschach which capitalize on ambiguity.)

Item forms using fixed response-categories are particularly open to criticism. The attitude-test pattern, where the subject marks a statement A, a, U, d, or D, according to his degree of agreement, is open to the following response sets: Acquiescence, or tendency to mark "A" and "a" more than "d" and "D"; evasiveness, tendency to mark "U"; and tendency to go to extremes, to mark "A" and "D" more than "a" and "d". Prob-

ably not all three of these sets will operate to a significant degree in any given test, but it is better to eliminate the sets at the outset than to spend effort later trying to measure the effect of the sets and root them out. Test designers generally have argued for retaining the five-point scale of judgment, or the more indefinite seven-point, ten-point, or even continuous scales. Such scales are open to marked individual differences in definition of the reference positions, with the more complex scale offering more chance for personal interpretation. The usual argument for the more finely divided scale of judgment on each attitude item is that it is more reliable and that subjects prefer it. If the latter advantage is significant, the finer scale may be retained and scored dichotomously. The argument that the finer scale gives more reliability is not a sound one, since this is precisely what we would expect if all of the added reliable variance were response-set variance and had no relation to beliefs about the attitude-object in question. There is no merit in enhancing test reliability unless validity is enhanced at least proportionately. It is an open question whether a finer scale of judgment gives either a more *valid* ranking of subjects according to belief, or (what we are beginning to recognize as even more important) scores more *saturated* with valid variance. With raters trained to interpret the scale uniformly, so that response-set variance is removed, the finer scale may be advantageous.

The writer therefore renews his earlier recommendation that the following forms of item be avoided in tests where high validity is more important than speed-of-test construction: true-false, like-indifferent-dislike, same-different, yes? no, agree-uncertain-disagree, and mark all correct answers. What does this leave? Foremost, it leaves the forced-choice or best-answer test. Our attempt to find a response set in the multiple-choice test was almost completely unsuccessful. A set was extracted, and that a set with little reliability, only when the test was applied to subjects for whom it was unreasonably difficult. Further studies of multiple-choice tests are still in order, but experience to date justifies the assumption that they are generally free from response sets. One confirmation of the argument that forced choices should be used comes from a

study by Owens (18). He found that substituting forced-choice for the "yes-no" response of the conventional neurotic inventory significantly reduced the number of false positives, i.e., it increased empirical validity. The forced choice has long been used successfully in many fields. Tests of mental ability now use it almost to the exclusion of other forms. Spelling, arithmetic, and grammar tests can certainly be cast in "recognize the right (or wrong) choice" form, rather than checklist forms and others open to response sets. Thurstone used it successfully in his paired-comparison approach to attitudes, and the same approach has long been found satisfactory in psychophysics. The Kuder interest test is well known, and Kuder has recently developed a new test of personality in the same forced-choice form. Paired comparisons may serve well in employee rating, and the Army has found the forced-choice valuable in obtaining officer ratings. Apparently forced-choice items can be used for nearly all purposes now served by the inadequate item forms.

Another important consideration is test difficulty, regardless of item form. The influence of response sets rises with difficulty, and therefore measurement of differences between students who find the test difficult is particularly invalid. This is, first, a reason for not using a test on subjects for whom it is quite difficult. Second, however, it suggests basing measurement on scales of adaptable difficulty. Thus, with the Kuhlmann-Anderson mental-test series, one selects the scales which have a difficulty appropriate for the subject, and if the first tests tried prove to be too difficult, the tester can move to an easier set of items to obtain more accurate measurement. Tests of this type, which are common in psychophysics, would be hard to use in group measurement; but experimental trial of such test designs is worth considering. If the *Seashore Pitch Test*, for example, were redesigned, one might have a preliminary section of twenty (?) items, ranging from very hard to very easy. This could be scored as soon as completed, and if the score were high, the subject would be given a difficult 50-item test (perhaps with all differences five cycles or two cycles). But a subject who performed near the chance level on the preliminary test would be given a final test of items with large differences (per-

haps 20 to 30 cycles). A set of several overlapping scales would be required, all standardized on the same group. Such a test could not test large groups inexpensively, but could be quite accurate in testing individuals.

(b) *Modification of directions.*—If, in any test, we expect a particular response set to arise, we can revise the directions to reduce the ambiguity of the situation. Another way of accomplishing the same end is to give students general training in test-wisness. For example, if they know that in most true-false tests about half the items are false, they will tend to avoid excessive acquiescence. If they know that the correction formula is based on chance, they will know that the odds are in their favor when they respond to items where they are uncertain.

It appears to the writer that, in most tests, subjects should be directed to answer all items, even though this tends to increase the random error variance. In many situations, this source of error is less damaging than the constant errors introduced by differences in tendency to guess, checking threshold, or diligence in searching for correct answers. Wesman (25) reports partial evidence that grammar items, where the subject marks each error he notices in given sentences, become more reliable when the subject is directed to mark every sentence-part "correct" or "incorrect," rather than just checking the "incorrects" (but evidence on validity is lacking).

Whisler (27) raised the question of response-habits in Thurstone-type attitude scales. He found that some subjects marked six or more items in a 22-item scale, and for them the reliability (parallel-test) of the attitude score was .89. But for the subjects who marked five or fewer items that they agreed with, the reliability was .62. Whisler thought that the subjects who checked more items were more careful in using the scale, or that their attitudes were more integrated. Hancock (8) followed Whisler with an experimental alteration of directions. First, he directed subjects to mark all the statements they accepted, then the five with which they most agreed, and, finally, the three of that five which they most strongly accepted. The shift of directions produced some alteration in scores. Generally, the standard deviation (in scale value) of scores increased when

fewer items were counted. For those with attitudes favorable to an occupation, the more items they checked, the closer their score was to the indifference position. Unfortunately, there is not enough evidence in the Hancock report to give a basis for selecting any particular number of checks as preferable. If the number of items checked affects mean, sigma, and reliability, there can be little justification for permitting the number to vary. It appears desirable to require every subject to mark a fixed number of alternatives, selecting the statements with which he most agrees. Limited experience with this procedure suggests that the subject should check around one-fourth of the statements.

(c) *Correction for response sets.*—When response sets are entering scores on a test, we may control or correct for the effect by special scoring keys. One widely used method is the control score. If a "response-set score" can be obtained, we may identify all cases with extreme response sets and drop such cases from the sample, admitting that measurement for them is invalid. The most familiar examples appear in the control scores of the Minnesota Multiphasic. Many other tests also permit us to derive such scores as bias or acquiescence, or number of items marked. In some tests it may be acceptable to report two scores for every subject; all the essential data in the hypothetical spelling test discussed earlier could be reported in one score "number right" and a second "number marked as incorrect." But simultaneous consideration of patterns of scores is awkward.

Humm has long used the No-Count as a control score on his Temperament Scale. A comment in the Supplemental Manual for that test is of interest:

It was observed that subjects whose scores in the Scale were at variance with the results of case studies by psychiatrists, psychologists, and social workers were found more often among those with an ultra-high or an ultra-low proportion of no-responses, than was the case where no-responses were in the middle ranges. Individuals who answer the questions of the scale with a high number of no-responses tend, consciously or unconsciously, to obscure their real temperaments. On the other hand, individuals with a low number of responses may exaggerate their temperamental characteristics.

Eliminating cases with extreme control-scores has the disadvantage of throwing out numerous subjects, but it is vastly better than treating the subjects as if the scores were valid. Sometimes a simple solution is to readminister the test with more careful directions, as Bennett and others illustrate (1). But more complex correction procedures are possible. In this, Humm and his co-workers were also pioneers.

Two procedures have been developed for cases where No-Counts are extreme. The first is the "profile score." For an initial sample of 181 cases, Humm had a criterion score on each component the test claimed to measure. The profile score is the best estimate of the criterion score from the uncorrected score and the No-Count. This procedure, regressing from an external criterion rather than merely partialling out No-Count in terms of the zero-order r between No-Count and raw component score, allows for the very reasonable assumption that part of the No-Count variance represents significant elements in personality.

The second correction, reserved only for cases where profile scores are inadequately revealing, yields the "regression score." This "stated the standard deviational distance of the given component score from the mode of scores in that component attained in scales showing the same No-Count. The regression score takes no account of validity. It does not, therefore, consider how well the Component Score measures the 'true' component strength." This, of course, partials out all the response-set portion of the score variance.

Humm and Humm (12) report that their procedures raise the validity of interpretations, for those papers where correction is required. Similar methods could no doubt be applied to other tests, and in the K-correction of the Multiphasic, a similar treatment is illustrated. Such refined statistical improvements are worth making only when one intends to treat a test quite seriously. It would scarcely be worthwhile to build a correction score for acquiescence into the Bernreuter test, in view of the many other bases for doubting its validity. But where great statistical labor in the form of factor analysis has already entered such a test as Guilford's series, application of a control score for response sets may be worth serious consideration.

Correction for response sets is a problem in suppressor variables (10, pp. 140-142). We wish to retain valid response-set variance (Type 3a), but we wish to remove from the score the variance of Type 3b and 2. If an independent estimate of the Type 3a variance, or of the combined undesirable variance, could be obtained by a pure measure of the response set itself, this estimate might be used as a suppressor variable.

Capitalizing on response-set variance. If response sets are thought of as possibly contributing to validity, one may weight the response sets in a way that maximizes their contribution. Cook and Leeds (3) correlate each possible response on an attitude scale for teachers with a criterion, and assign positive or negative scoring weights accordingly. One item is as follows, where the numbers in parentheses are weights:

	1	2	3	4	5
It is some-	Strongly	Agree	Un-	Disagree	Strongly
times neces-	agree		decided		disagree
sary to break	(0)	(4)	(-1)	(4)	(-1)
promises to					
children.					

The criterion used was a dependable estimate of the ability of teachers to establish rapport with children, which the scale was supposed to predict. It will be noted that the scoring weights are "illogical," since there can be no stronger response to "It is *sometimes* necessary. . . ." than to disagree (response 4), which amounts to saying "It is *never* necessary." The weights for responses 4 and 5 reflect the difference in response set (not in logically considered opinion) between teachers in the superior and inferior criterion groups. The defense of the Cook-Leeds procedure, and the comparable method used in Strong's Interest Blank, is that it yields considerable validity. The limitation is that invalid variance (Types 2 and 3b) is weighted just like valid variance. A particular "good" teacher who has a set to respond very emphatically will be penalized by the weights. The majority of "good" teachers, who avoid extreme responses, will be reliably discriminated by the key. One difficulty with the sheer empiricism represented here is that the weights serve their practical purpose but give little insight into the nature of the variables tested. The only basis for extending or improving the test is trial-and-error, developing

many more items of all sorts and trying them to see how the weights come out.

Sometimes, instead of employing correction scores to refine the total test score, one may modify the original test scores. Thus Flanagan (23, p. 9) suggests scoring Rights and Wrongs separately, and using each score in the multiple-correlation when trying to predict a criterion. This procedure permits one to weight "carefulness" variance separately from "ability" variance. Work with true-false tests suggests that scores Rights-on-True-Items and Rights-on-False-Items will have different validity and may be assigned different weights in the predictor score (5). Probably this notion could be extended further, in empirical prediction.

Summary

This paper summarizes extensive evidence demonstrating that such response sets as bias in favor of a particular alternative, tendency to guess, working for speed rather than accuracy, and the like, operate in conventional objective tests. Not only are such sets widespread, but they reduce the validity of test scores. The response set can be altered readily by alteration of the directions or by coaching. Some studies show that response sets are somewhat correlated from one test to another (but not if the tests differ greatly in content), and that they are correlated with important external variables. While response-set variance may under certain circumstances enhance logical and empirical validity, it appears that its general effect is to reduce the saturation of the test and to limit its possible validity.

The following recommendations for practice, most of which were previously suggested, are reinforced by the present findings:

1. Response sets should be avoided with the occasional exception of some tests measuring carefulness or other personality traits which are psychologically similar to response sets.
2. The forced-choice, paired-comparison, or "do-guess" multiple-choice test should be given preference over other forms of test item.
3. When a form of item is used in which response sets are possible,

- a) Directions should be worded so as to reduce ambiguity and to force every student to respond with the same set.
 - b) The test should not be given to a group of students for whom it is quite difficult.
 - c) A response-set score should be obtained, and used to identify subjects whose scores are probably invalid.
4. Where response sets are present, attempts should be made to correct for or to capitalize on the response set by an appropriate empirical procedure.

In view of the overwhelming evidence that many common item forms invite response sets, and in view of the probability that these sets interfere with accurate measurement, it will rarely be wise to build new tests around item forms such as A-U-D, Yes-No-?, and "check all correct answers." It is to be hoped that the tests forthcoming in the future will be designed to increase their saturation with the factors the test is seeking to measure.

REFERENCES

1. Bennett, George K., Seashore, Harold G. and Wesman, Alexander G. *Differential Aptitude Tests, Manual*. New York: Psychological Corporation, 1947.
2. Brotherton, D. A., Read, J. M. and Pratt, K. C. "Indeterminate Number Concepts: II. Application by Children to Determinate Number Groups." *Journal of Genetic Psychology*, LXXIII (1948), 209-236.
3. Cook, Walter, W. and Leeds, Carroll H. "Measuring Teacher Personality." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, VII (1947), 399-410.
4. Cronbach, L. J. "Response Sets and Test Validity." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, VI (1946), 475-494.
5. Cronbach, L. J. "Studies of Acquiescence as a Factor in the True-False Test." *Journal of Educational Psychology* XXXIII (1942), 401-415.
6. Goodfellow, Louis D. "The Human Element in Probability." *Journal of General Psychology*, XXXIII (1940), 201-205.
7. Guilford, J. P. (ed.) *Printed Classification Tests*. AAF Aviation Psychology Program Research Reports, No. 5. Washington, D. C.: Government Printing Office, 1947.
8. Hancock, John W. "An Experimental Study of Limiting Response on Attitude Scales." In H. H. Remmers (ed.), *Further Studies in Attitudes*, Series III. *Studies in Higher Education*, XXXIV. Lafayette, Ind.: Purdue University, 1938. Pp. 142-148.

9. Harlow, H. F. "The Formation of Learning Sets." *Psychological Review*, LVI (1949), 51-65.
10. Horst, Paul. *The Prediction of Personal Adjustment*. New York: Social Science Research Council, 1941.
11. Humm, Doncaster, G. and Wadsworth, Guy, Jr. *The Interpretation of the Humm-Wadsworth Temperament Scale*. Los Angeles: D. G. Humm, 1943.
12. Humm, Doncaster G. and Humm, Kathryn A. "Compensations for Subjects' Response-Bias in a Measure of Temperament." *American Psychologist*, II (1947), 305.
13. Johnston, Aaron Montgomery. "The Relationship of Various Factors to Autocratic and Democratic Classroom Practices." Unpublished doctoral dissertation, University of Chicago, 1948.
14. Kahn, D. F. and Hadley, J. M. "Factors Related to Life Insurance Selling." *Journal of Applied Psychology*, XXXIII (1949), 132-140.
15. Lorge, I. "Gen-like: Halo or Reality?" *Psychological Bulletin*, XXXIV (1937), 545-546.
16. Mathews, C. O. "The Effect of the Order of Printed Response on an Interest Questionnaire." *Journal of Educational Psychology*, XX (1929), 128-134.
17. Mersman, Ivo. "Personality Traits as Related to Vocational Choice." Unpublished masters' thesis, University of Chicago, 1948.
18. Owens, W. A. "Item Form and 'False-Positive' Responses on a Neurotic Inventory." *Journal of Clinical Psychology*, III (1947), 264-269.
19. Philip, B. R. "Generalization and Central Tendency in the Discrimination of a Series of Stimuli." *Canadian Journal of Psychology*, I (1947), 196-204.
20. Rubin, Harry K. "A Constant Error in the Seashore Test of Pitch Discrimination." Unpublished masters' thesis, University of Wisconsin, 1940.
21. Singer, William B. and Young, Paul T. "Studies in Affective Reaction: III. The Specificity of Affective Reactions." *Journal of General Psychology*, XXIV (1941), 327-341.
22. Thorndike, R. L. "Critical Note on the Pressey Interest-Attitudes Test." *Journal of Applied Psychology*, XXII (1938), 657-658.
23. Vaughn, K. W. (ed.) "National Projects in Educational Measurement." *American Council on Education Studies*, Series I, No. 28, (1947), pp. 8-12.
24. Vernon, P. E. "Classifying High Grade Occupational Interests." *Journal of Abnormal and Social Psychology*, XLIV, (1949), 85-96.
25. Wesman, Alexander, G. "Active versus Blank Responses to Multiple-Choice Items." *Journal of Educational Psychology*, XXXVIII (1947), 89-95.

26. Wesman, Alexander G. "The Usefulness of Correctly Spelled Words in a Spelling Test." *Journal of Educational Psychology*, XXXVII (1947), 242-246.
27. Whisler, L. D. "'Reliability' of Scores on Attitude Scales as Related to Scoring Method." in H. H. Remmers, (ed.), *Further Studies in Attitudes*, Series III. *Studies in Higher Education*, XXXIV. Lafayette, Ind.: Purdue University, 1938. Pp. 126-129.
28. Wyatt, Ruth F. "Improvability of Pitch Discrimination." *Psychological Monographs*, LVIII (1945), No. 267.

CLIENT ACCEPTANCE OF SELF-INFORMATION IN COUNSELING

ROBERT B. KAMM

Drake University

and

C GILBERT WRENN

University of Minnesota

THE counseling process is generally recognized today as a professional psychological function. Writers on the subject agree that the counseling experience is a dynamic relationship between two people - an ever-changing relationship to which many variables contribute. This concept has emerged as a result of three somewhat varied, yet related, types of research on the counseling process: studies of evaluation, studies of counseling methodology, and studies of factors operative within the counseling interview. The present research study is classified in the last group in that it is a consideration of factors at work within the interview situation.

Research has shown that certain students seem to benefit from the counseling process. On the other hand, other students apparently do not benefit from this experience. Some students appear to follow counselor suggestions and to accept information more readily than do others. The question arises: Within which interview situations do clients tend to accept information presented, and within which do they tend not to accept the data? Further, in what ways, if any, do students who accept the information differ from those who do not? Also, what types of information tend to be accepted, and what types tend not to be?

In the present study "acceptance" is defined as favorable reception by the client of information presented to him, as demonstrated by (a) what the client says and (b) what the client does. "Information" includes all data presented by the counselor, whether they be in the form of advice, suggestion,

emphasis, recommendation, interpretation, request or explanation. The type of interview in the present study is limited to educational-vocational planning interviews.

Methodology of Study

Utilizing one trained, experienced counselor, complete phonographic recordings were made of forty educational-vocational planning interviews. The clients used were University of Minnesota first-quarter General College freshman men who voluntarily sought counseling. They were typical of the General College population with regard to academic ability and vocational interests.

Just prior to the actual recording, each of the clients completed an "immediate pre-interview" form of inquiry pertaining to his educational and vocational plans. Within several days following the recorded interview, an interview during which the client's academic ability, interest, and aptitudes had been discussed with him, with suggestions and recommendations made by the counselor, each client completed an "immediate post-interview" form of inquiry. In addition, the counselor after each interview indicated on a check list his judgment with regard to the emotional states of the client and counselor and the degree of rapport achieved in the interview.

At one month and again at four months after the recorded interview, the investigator interviewed each of the clients in an effort to gain additional evidence for and against acceptance on the part of each client. Following this, all of the pre-interview and post-interview data were summarized for each case and presented to a team of three judges who, working independently, decided in which cases "acceptance" had occurred or had not occurred. A composite of the judges' decisions was made in order to categorize the cases as acceptance or non-acceptance.

In the meantime written transcriptions had been made of each of the forty recorded interviews. Following this, each of the client and counselor responses, numbering 12,238, were categorized into one of twenty-two categories.

For the classification of counselor responses, Seeman's nine categories were used (4). These are: (a) counselor questions

dealing with content or factual data; (b) counselor questions concerned with the attitudes and motivations of the client; (c) counselor responses to content; (d) counselor responses to feelings, attitudes and motivations of the client; (e) counselor interpretations and opinions concerning content; (f) counselor interpretations and opinions concerning client attitudes, feelings and motivations; (g) suggestions, advice and counselor decisions on courses of action for the client; (h) suggestions, advice and counselor decisions concerning client attitudes and feelings; (i) information given by the counselor.

In addition to Seeman's nine categories, two additional counselor-response categories were used in the present study. These were: (a) unclassified and (b) simple agreement: ("Yes," "uh-huh").

In the categorization of client responses, Snyder's eight general categories for client content and three general categories for client-feeling responses were employed in the study (5). *The client-content categories* are: (a) problem; (b) asking for information; (c) disagreements; (d) answering questions; (e) agreement; (f) insight; (g) planning; and (h) miscellaneous.

The client-feeling categories include: (a) positive attitudes (statements which reveal approval and acceptance of the client himself, the counselor or the counseling process or other persons, objects or situations); (b) negative attitudes (statements which reveal disapproval or rejection of the client himself, the counselor or the counseling process, or other persons, objects or situations); (c) ambivalent attitudes.

Pertinent personal data such as previous work experience, education and home background, as well as academic-aptitude test scores and interest and personality inventory results, were also gathered for each of the forty cases.

Findings of the Study

The composite rating of the judges showed that in twenty-six of the forty cases the clients either "definitely" or "for the most part" accepted information presented in the interview. The other fourteen cases were divided among the "indecisive" cases and the "definitely" and "for the most part" non-acceptance cases. The heavy weighting of acceptance cases may be

attributed in part to the fact that the clients came voluntarily for help with their problems.

The most important findings of this investigation pertinent to the dynamics of acceptance are the following:

1. Client acceptance of information presented occurs most often in those situations in which *both client and counselor are completely relaxed*. When either of the two, or both, are not relaxed, acceptance is less likely to occur.
2. *Acceptance is directly related to "positive attitude" as expressed by clients during the interview*. Acceptance, on the other hand, is inversely related to both negative and ambivalent attitudes, as expressed by clients during the interview.
3. Acceptance is directly related to a "readiness" for counseling help. Merely having a "felt need" on the part of the client does not necessarily mean that acceptance of information, pertaining to that need, will occur. A *readiness to act with regard to a felt need* appears to be the crucial factor with regard to acceptance.
4. Information which is *directly related to the client's own immediate problem* tends to be accepted.
5. Information which is *not in opposition to client self-concept* tends to be accepted. Further, information which shows the client to be like others of his group tends to be accepted whereas information which shows him to be deviate tends not to be accepted.

Less crucial findings of the study are:

1. The counselor used in the present study did not differ significantly in his counseling approach for the acceptance and the non-acceptance groups. It was found that, as the interview progressed, he (a) asked fewer questions, (b) gave more suggestions and directions, and (c) showed less simple agreement. He showed an increase in information-giving from the initial one-third of the interview to the middle one-third and then a decrease from the middle to the final one-third. He made little use of feeling-responses.
2. Although the counselor's approach in the interview situation does not vary significantly in the present study,

some clients accept information presented, whereas others do not. This suggests the operation of factors *other than the counselor's approach* in the determination of client acceptance.

3. Both acceptance and non-acceptance of information can occur in situations in which the client-counselor relationship is friendly. Also, when an apathetic relationship is experienced, either acceptance or non-acceptance can occur.
4. There appears to be a positive relationship between non-acceptance and the achievement of only a "surface" understanding of the problem by the client and the counselor, as indicated by the counselor's rating of the interview.
5. Acceptance does not appear to be related to client use of such categorized responses during the interview as (a) statement of the problem, (b) answering of counselor questions, (c) indications of insight gained, (d) indications of plans, and, (e) unrelated client discussion. The data suggest (although the findings between acceptance and non-acceptance groups are not statistically significant) that acceptance may be related to client agreement and inversely related to client disagreement, as shown by client responses during the interview. Acceptance may also be inversely related to the asking of factual questions during the interview.
6. For both the acceptance and the non-acceptance cases as the interviews progressed, there was (a) a decrease in client statement of the problem, (b) an increase, followed by a tapering off, in the asking of questions by the client, (c) a decrease in client answering of counselor questions, (d) an increase in client agreement, and (f) an increase in client statements pertaining to plans. The non-acceptance group showed an increase in unrelated statements, whereas the acceptance group was constant with regard to unrelated data.
7. The non-acceptance cases, like the acceptance cases, showed an increase in the expression of positive feelings as the interview progressed. The level of expression of

positive feelings, however, was significantly lower throughout the interview for the non-acceptance group. The two groups likewise showed parallel patterns of decrease of negativism and ambivalence.

8. Acceptance appears to be unrelated to the factors of (a) the length of the interview, (b) the time of the day of the interview, and (c) the proportion of time which the client speaks during the course of the interview.
9. Acceptance is unrelated to (a) academic aptitude, (b) particular measured personality patterns, (c) social status of the client's home, (d) veteran status, (e) marital status, (f) part-time work status while in college or (g) the factor of previous client-counseling contacts.
10. For those judged to have "definitely" accepted information presented, there appears to be a direct relationship between acceptance and good first-quarter academic achievement.
11. With regard to vocational interest patterns, acceptance appears to be related to the presence of interest profiles which contain all three types of interest patterns: primary, secondary, and tertiary. Except for this finding, acceptance is unrelated to any particular vocational interest pattern.
12. Different kinds of information are accepted equally well by the acceptance and non-acceptance groups, with one possible exception. Information which involves an altering of previously made client plans tends to be more often accepted by the group defined as the "acceptance" group.

Conclusions and Implications for Counseling

Conclusions obtained from the present findings and implications for counseling follow:

1. The importance of certain psychological factors in the acceptance of information has been noted. *The most conclusive of all the findings, perhaps, is that acceptance is related to client feeling, particularly feeling or attitude toward self.* The importance of an emotionally relaxed client-counselor relationship has been shown. The factor of "readiness" has been indi-

cated. Further, it has been pointed out that information which is directly related to the client's own immediate needs is likely to be accepted, as is information which does not oppose or injure the client self-concept.

The counselor must recognize the presence of positive, negative, and ambivalent attitudes of the client. If the client shows a predominance of negativism and/or ambivalence, it may be necessary that the counselor structure the counseling process in such a manner that there would be a series of "preparation for educational-vocational planning" contacts, devoted to the development of proper client sets and attitudes. Once this is done, acceptance of information pertaining to educational-vocational planning might take place more readily.

On the other hand, if the client demonstrates a warmth toward the interview, toward the counselor, toward himself, as well as toward others, the planning interview can proceed and the counselor may feel reasonably certain, other factors being equal, that acceptance of information will occur.

The finding that there is an increase in positive expression and decreases in negativism and ambivalence for the non-acceptance cases, as the interview progresses, poses an interesting problem. In the first place, this finding should be indicative to the counselor that all clients who "warm up" during the interview will not necessarily accept. More important, however, this demonstrated rise in positive feelings may be interpreted as an encouraging sign—a sign that may be indicative of acceptance at some later time, providing the client is given proper orientation and preparation for the educational-vocational planning session.

On the other hand this warming-up may be merely an expression of a pleasant social convention. In our culture all of us are taught to be as agreeable as possible, to put our best social face forward. Hathaway (3) has called this the "hello-goodbye" convention, this tendency to be pleasant and to express formal gratitude at the end of the interview. He warns against utilizing such expressions of goodwill in the interview as measures of the effectiveness of the interview. Hence this rise in positive feelings *may be* only a measure of the social graciousness of the client.

Closely allied is the factor of "readiness." Negativism and ambivalence may actually be indicative of a lack of readiness in some cases. The counselor must determine whether or not the client is ready for educational-vocational planning. If he is not ready, perhaps there will need to be "preparation for planning" contacts, as previously mentioned. In this connection one should not forget Butler's differentiation between the adjustment and distributive phases of counseling (2). He contends that the distributive phase (Kefauver's term, the use of "planning phase" might be even more appropriate) should not be entered upon until adjustment to the present and to himself has been assured. Thus the "preparation for planning" spoken of here may mean adjustment counseling using permissive methods of treatment. The "readiness to act" mentioned earlier may merely mean a *lack* of preoccupation with areas of self-regard other than those associated with the planning at hand.

The establishment of an emotionally relaxed relationship between client and counselor is, apparently, necessary before information will be accepted. The importance of good rapport in the interview relationship has long been recognized. The importance of an emotionally relaxed state as a contributor to good rapport is specifically noted here. The counselor is obligated to establish, insofar as possible, such a relationship.

The counselor must recognize what needs are most immediate and most pertinent to the client. *That which the client recognizes as real, not what the counselor sees, is most important.* Accordingly, if acceptance is to occur, it is necessary to start at the level where the client operates at the moment. The counselor must use techniques designed to assist in the development of the client to a more realistic awareness of himself. Here is introduced again the factor of readiness or of need for self-adjustment, showing how the factors related to acceptance are not discrete but intertwined.

The importance of starting at the level of thinking of the client is further given support by the present finding that information not in line with previous client plans tends to be rejected. The counselor must be aware of these previous plans, goals and objectives of the client. It is necessary for him to

recognize them and to take them into consideration if acceptance is to occur.

2. The lack of relationship between acceptance and the content of client response during the interview serves in a sense to point up and to give emphasis to the importance of client feeling. It appears that the counselor will do well to pay less attention to the content of what the client says and more to how the client feels.

3. With regard to acceptance, such traditionally stressed factors as academic aptitude, personality patterns, vocational interest patterns, home background and previous counseling contacts show little or no relationship to acceptance. These data, useful as they are in some situations, do not seem to be crucial insofar as acceptance is concerned. The acceptance of information presented apparently can occur in spite of low academic ability or a poor home background. Likewise, such factors as the length of the interview, the time of day of the interview, and the proportion of time in which the client speaks during the interview may not deserve the attention they are sometimes given, at least as far as acceptance is concerned.

4. The client with his needs, his wants and desires, his attitudes and feelings is the basic determiner of whether or not acceptance occurs. The data suggest that the client himself is more important than the interview situation itself or the type of information presented.

5. Certain individuals may benefit little, if any, from a particular counseling contact. In the present study with a group of college freshmen (typical of General College freshmen in general), there appears to be little reason for admitting that many of these students would *never* accept information presented during an educational-vocational planning interview. The evidence seems to indicate that all of the clients in the present group, with further attention, might develop to a state of acceptance. The possibility needs to be explored that there would be a greater acceptance of test information if test selection were made by the clients in the manner suggested by Bordin and Bixler (1). Theoretically such client-chosen tests would be in personality areas where there is adequate "readiness" for acceptance of results. Any attempt to assist the client

toward a realistic self-acceptance must always start at the level of the client and must be developed at a pace agreeable to the client. Such a technique assumes that the counselor has sufficient insight to recognize underlying client problems and to see them in the total picture of the client's existence.

A Proposal For Further Research

The present study might be regarded as a "pilot study," inasmuch as it was limited in scope and was a pioneering venture with regard to the methodology of the problem. For practical considerations the size of the sample was limited. In order to permit generalization from the small sample, the sample was restricted to one stratum of the college population, thereby securing a more homogeneous group. To limit the variables operative within the interview situation, only one counselor was used. These limitations were deemed necessary for the present study. A similar study should be carried out in which the following conditions might be observed.

1. A larger sample, representative of the total college population, should be utilized.
2. Several trained counselors should be employed to do the counseling. The counselors used should be known to vary from the more "non-directive" to the "directive" approach with regard to counseling philosophy and methodology. They might include those avowedly "eclectic" or those psychoanalytic in orientation.
3. Recorded data should include the entire series of contacts with each case rather than only one interview.
4. Problems not only pertaining to educational-vocational planning needs, but to other problem areas as well, should be studied.
5. Careful investigation of pre-interview behavior and rigorous observation of post-interview behavior should be done.

If the above suggestions were followed, a pool of data would be available which would provide many answers concerning the dynamics of the counseling process. Such a pool would provide data for any degree of intensity or extensiveness that would be desired. Acceptance of data pertaining to emotional and personality problems might be investigated. Detailed analysis of client sets which are brought to the interview could be made. Likewise, such other psychological aspects of the

interview as client motivation and the interaction of two personalities participating in the interviews could be studied. Studies of counselor methodology and careful analysis of segments of the interview could be made. The agency or institution which is willing to provide sufficient backing for such an enterprise will step into a position of leadership in counseling research.

REFERENCES

1. Bordin, Edward S. and Bixler, Ray H. "Test Selection: A Process of Counseling." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, VI (1946), 361-373.
2. BUTLER, JOHN M. "On the Role of Directive and Non-Directive Techniques in the Counseling Process." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, VIII (1948), 201-209.
3. HATHAWAY, STARKE R. "Some Considerations Relative to Non-Directive Counseling as Therapy." *Journal of Clinical Psychology*, III (1948), 226-231.
4. SEEMAN, JULIUS. "A Study of Preliminary Interview Methods in Vocational Counseling and Client Reactions to Counseling." Unpublished Ph.D. thesis, University of Minnesota, 1948.
5. SNYDER, WILLIAM U. "An Investigation of the Nature of Non-Directive Psychotherapy." *Journal of General Psychology*, XXXIII (1945), 193-223.

THE CONCEPTS OF RELIABILITY AND HOMOGENEITY

C. H. COOMBS¹
University of Michigan

I. *Introduction*

THE literature of test theory is replete with articles on the computation and interpretation of indices of reliability. In them one finds surprisingly little common agreement or even mutual understanding (6). In more recent years the concept of homogeneity, with its indices, has been added, with the result that the confusion has increased. We shall make no effort in this paper to review and summarize this literature but shall attempt to do three things:

- (1) point out what we regard as the fundamental sources of this confusion;
- (2) provide a theoretical foundation on the basis of which this confusion might be resolved;
- (3) point out the further steps that must be taken to develop the theory and practice of mental testing.

II. *Sources of Present Confusion*

There are two fundamental sources² of confusion in present test theory: one is the assumptions by means of which we arrive at an interval scale (3), and the second is the identification of

¹ This paper is an extension to the area of mental testing of some of the ideas contained in a chapter in a general theory of psychological scaling developed in 1948-1949 under the auspices of the Rand Corporation and while in residence in the Department and the Laboratory of Social Relations, Harvard University. While the author carries the responsibility for the ideas contained herein, their development would not have been possible without the criticism and stimulation of Samuel A. Stouffer, C. Frederick Mosteller, Paul Lazarsfeld, and Benjamin W. White in a joint seminar during that year. Development of the theory before and after the sojourn at Harvard was made possible by the support of the Bureau of Psychological Services, Institute for Human Adjustment, Horace H. Rackham School of Graduate Studies, University of Michigan. A version of these ideas was presented in a 1949 APA symposium on Test Homogeneity and Test Validity.

² A Complete discussion of the fundamental difficulties in present test theory is to be found in Thomas (5).

our statistical indices with the concepts they are presumed to measure. These two basic difficulties are intimately related and are both associated with our attempt to model psychological measurement on physical measurement. Let us discuss them briefly, in turn.

Consider the manner in which data are obtained in the area of mental testing: The method used is the method of single stimuli, in which there is one response from each individual to each stimulus. These responses comprise our basic data, and consist of two piles of items for each individual. One pile has the items which the individual passed and the other pile those items which he failed. Note that there is no information in the data for a given individual pertaining to (1) how well he passed one item compared with another, or (2) how badly he failed one item compared with another, or (3) finally, how badly he failed one item compared with how well he passed another. The only way to obtain metric relations in data collected by the method of single stimuli is to put the information in the data by means of a priori statistical assumptions concerning, for example, the shape of the distribution function of the abilities of the individuals on the attribute in question. A normal distribution is usually what is assumed in test theory but even this is not applied in a thoroughgoing fashion.

To carry out the assumption fully (1) the percentage passing each item should be corrected for chance, then (2) converted to a sigma score, and (3) items at equal intervals on this sigma scale should be selected for a final form. This procedure is usually not rigorously adhered to because, in the first place, it makes little practical difference, in many instances, if the items are not precisely distributed in a discrete rectangular distribution on this sigma scale. But there is another reason why it is not insisted that this procedure should be rigorously adhered to, and that is because the assumptions which lead to a unit of measurement implicitly require the further assumption of perfect homogeneity. The distrust of the procedure is supported by the fact that the assumption of perfect homogeneity can usually, if not always, be shown to be violated, even in such crude data as that collected by the method of single stimuli. Unfortunately, to many this is simply regarded as one of the

sources of error variance and not as a fundamental theoretical obstruction.

Thus, in the method of single stimuli as applied to mental testing we create an interval scale without any built-in or inherent test of its validity. Having such a scale, then, it is permissible to use certain properties of numbers, and we have available a variety of statistical procedures for the analysis of behavior. We must, of course, allow for error variance, much of which we have put there ourselves in assuming an interval scale, and, consequently, a statistical theory of error becomes necessary and plays a dominant role in test theory. This, then, is one major source of difficulty in the area of tests and measurements but, important as it is, it is not as fundamental as the second source. The difficulty arising from assumptions leading to an interval scale is of significance primarily to the empirical aspect of psychological testing rather than to the theoretical aspect.

The second source of difficulty, which we consider to be of prime theoretical significance, has, however, arisen from the use of an interval scale. Basically, this second source of confusion is the fact that we have had no fundamental *psychological* rationale underlying our concepts in test theory. Rather, we find an easy road to the concepts of test score, difficulty of an item, reliability and homogeneity via statistical definitions of indices dependent upon the existence of an interval scale. We set up these statistical indices based on operational procedures, then give names to them and act as if they have certain obvious psychological meanings. We have gained readily obtainable empirical indices but have paid for them in psychological ambiguity and imprecise meanings and interpretations. While relatively easy to compute and apparently readily susceptible to empirical study, an invalid assumption of an interval scale would vitiate even their numerical precision. Thus, we have not one but many indices of reliability, each determined in a different way, and hence each implying a different meaning. We do not have, independently, a quantitative definition of the concept of reliability, psychologically derived, with a unique interpretation. We have a variety of meanings for the concept of reliability, depending upon the index used. It is our thesis

that the concept of reliability should have a unique psychological meaning quantitatively defined, and the various indices should then be regarded as different kinds of approximations to the concept. The challenge, then, would be to the experimenter to devise indices which are better measures of the concept.

III. *A Psychological Rationale for the Concepts of Reliability and Homogeneity*

The Fundamental Equation.—We shall now attempt to sketch a theoretical psychological foundation for the derivation of quantitative definitions of certain concepts of test theory.

Consider the concept of the difficulty of an item. We all have intuitive notions as to what the psychological meaning of the difficulty of an item is. It means how hard it is for some one to pass it. But we identify the difficulty of an item with the percentage of people passing it. We thus have a number to represent the difficulty of an item which is the same number for all the people in the sample. Yet we know that for some people the item was so easy that they passed it, and for others it was so difficult that they failed it. It is apparent that we must have a definition of the difficulty of an item which will permit different values for different people. Of course, such a definition could still permit an *average* difficulty corresponding in principle to the conventional definition.

In order to develop a psychological rationale for the difficulty of an item let us consider an arithmetic problem. Let this arithmetic problem require that an individual know how to perform certain operations. The problem might involve addition and subtraction, the use of log tables, and a certain amount of reasoning. Its solution requires a collection of abilities, each to a certain degree and combined in a certain way. We may, for the sake of simplicity in discussion, lump this particular combination of abilities and call it a single ability. The problem then requires that every individual possess at least a certain amount of this ability in order to solve it. We shall call the quantity of an ability required for the solution of a problem the \mathcal{Q} value of that problem or that item.

Shall we regard this \mathcal{Q} value of an item as its difficulty? We

might, if we wish, so define the difficulty of the item. But this is not psychologically satisfying, because if we ask individuals how difficult an item is, some will say that it is easy and some will say it is difficult. How can the item have one \mathcal{Q} value and yet give rise to all this disagreement about its difficulty? Obviously it must be because these different individuals are making their judgments from different points of view. A mathematics major says it is easy; a grammar school student says it is hard. The point of view depends on the amount of this particular ability the person has. Of the particular ability demanded by the item, the amount possessed by an individual will be designated his C value, representing his capacity.

We have now a hypothetical continuum on which is a \mathcal{Q} value representing the amount of an ability required by the item from any individual to whom it is administered, and we have also a C value on this same continuum for each individual who attempts the item. How, then, shall we represent the degree of difficulty that this item has for a particular individual? This might be done in a number of ways. We have chosen to use the ratio of \mathcal{Q} to C to represent the psychological value or difficulty of this item for that individual and have called this ratio P , and thus we have the simple equation:

$$(1) \qquad \mathcal{Q} = PC$$

Obviously, the greater an individual's capacity the smaller proportion of that capacity is required or exercised in solving the problem and the easier it appears to him.

Each time (h) an individual (i) responds to a stimulus (j) here is a set of values which satisfy $\mathcal{Q}_{hij} = P_{hi}C_{hi}$. The most frequent objectives of psychological measurement are to determine something about the \mathcal{Q} values of each member of a set of stimuli and the C values of each member of a group of individuals.

But *note*, and this is significant to our later problem of metric, we do not observe \mathcal{Q} values and C values. Instead, what we observe are the P values. Thus, if an individual passes an item, we know that on that particular ability the individual's capacity¹, C_{ij} , was greater than the quantity², \mathcal{Q}_{ij} , required to pass the item and hence the P_{ij} value was less than one. In

¹ The subscript h is *one* here.

the method of single stimuli, which is the method most used in mental testing, we can divide the items into two categories for each individual, those whose P values were less than one for him, and those whose P values were greater than one. From such data on several individuals we want to extract what information they contain about Q and C values. If we refuse to make the assumptions which lead to an interval scale, exhaustive analysis of these data would yield, at best⁴, the *order* of the stimuli, (the Q values) and the *order* of the people (their C values).

We might digress for a moment to point out that with other methods of collecting data, such as the method of rank order, the method of paired comparisons, and the method of triads, we are able to collect, successively, much more information about the P values of stimuli for each individual and hence learn more about Q values and C values than we do from the method of single stimuli used in mental testing. Curiously enough it appears that we are going to be able to go further, with fewer assumptions, in the area of so-called qualitative attributes than in the area of mental testing.

The Variance of an Individual's Score: Imagine now that we have a stimulus or test item and a group of individuals who respond to it. Each individual's response to the item provides a P value. Of course we do not know the exact magnitude of a P value, we know only whether it is less than one or greater than one, that is, whether the individual passed or failed the item. But this is a limitation of this method of collecting data. Let us imagine that we had a method which would give us the exact P values. There would be, then, a distribution of P values for the stimulus. This distribution represents the distribution of difficulties which the item has for the individuals in the group.

Each individual has one of the P values in this distribution. Let us imagine that we could again administer this item to this same group of individuals *independently*⁵ of its previous ad-

⁴We have avoided the complication introduced by the true-false and multiple-choice type of item in which an individual may get an item right by pure chance. There is no need for this complication from the point of view of constructing a theory.

⁵The conditions necessary are that Q_{hi} be constant over h and i and the C_{hi} be constant over h and j . For purposes of future generalization these constitute an extreme of class I conditions (1).

⁶Experimental independence.

ministration. Then, once again, each individual would have a P value for this item. Would the successive P values of an individual for the one stimulus be identical, even if the successive administrations were independent? This is a question of whether or not P_{hij} is constant over h for a given i and j and can only be answered by experiment. It might well be that in the case of one attribute, say arithmetic, these successive P values would be almost constant for any given individual, whereas in the case of another attribute, say the aesthetic merit of a painting, the P values might be greatly variable. In this latter case we would expect the P values to be variable if the individual was not too clear as to just what he meant by aesthetic merit and hence used different criteria in successive evaluations of the painting. Thus, if the continuum is intrinsically different at different times, both the Q values of the stimulus and the C values of the individual would be variable for the same *nominal* trait, like aesthetic merit, because the exact composition of the trait was variable.

We have conceived, now, of each individual in a group having responded a number of times to a stimulus and, hence, for each individual, i , there is a distribution of P_{hij} values for the stimulus j . Let us now do the same thing for more stimuli, and imagine that there is for every individual a small distribution of his P values for each stimulus within the total distribution of all individuals' P values for each stimulus. The notation used is as follows:

$h = 1, 2, \dots, t$, (the number of times an individual responds to a stimulus)

$i = 1, 2, \dots, N$, (the number of individuals)

$j = 1, 2, \dots, n$, (the number of stimuli)

$$P_{ij} = \frac{1}{t} \sum_h P_{hij}$$

$$P_i = \frac{1}{nt} \sum_j \sum_h P_{hij}$$

$$P_j = \frac{1}{Nt} \sum_i \sum_h P_{hij}$$

$$\bar{P} = \frac{1}{Nnt} \sum_i \sum_j \sum_h P_{hij}$$

We are now in a position to define the status score, S_i (2), of an individual as follows:

$$(2) \quad S_i = \frac{1}{m} \sum_j \sum_k (P_{ij} - P_{kij})$$

or

$$(3) \quad S_i = \bar{P} - P_i$$

To put the status score of an individual in words, it is defined as the average difficulty of all the items for all individuals minus the average difficulty of all the items for him alone. Thus, we have made the score of the individual dependent upon the composition of the group of individuals of which he is a member. On this scale the average individual has a score of zero, and the better the individual the higher his score, since the easier the items are for an individual the smaller the proportion of his capacity is required to pass them and the larger would be S_i . Individuals below average would have negative status scores.

Inasmuch as, in principle, an individual has a score, an S_i , on every item every time he takes it, let us consider the composition of the variance of all these "scores" that get averaged together for a total score. If we designate by V_i the total variance of an individual, we have

$$(4) \quad V_i = \frac{1}{m} \sum_j \sum_k (P_{ij} - P_{kij})^2 - S_i^2$$

By adding and subtracting P_{ij} inside the parentheses, expanding and collecting terms, the expression for V_i becomes:

$$(5) \quad V_i = \frac{1}{m} \sum_j \sum_k (P_{ij} - P_{kij})^2 + \frac{1}{n} \sum_j (P_{ij} - P_{.j})^2 - \left[\frac{1}{n} \sum_j (P_{ij} - P_{.j}) \right]^2$$

Making the following definitions,

$$(6) \quad D_i^2 = \frac{1}{m} \sum_j \sum_k (P_{ij} - P_{kij})^2$$

$$(7) \quad T_i^2 = \frac{1}{n} \sum_j (P_{ij} - P_{.j})^2 - \left[\frac{1}{n} \sum_j (P_{ij} - P_{.j}) \right]^2$$

we have

$$(8) \quad V_i = D_i^2 + T_i^2$$

and V_i is seen to have two components. These two components, D_i and T_i , are of psychological significance. The first component, D_i , we call the individual's dispersion score and it represents the variability within an individual in repeatedly responding (independently) to the same stimulus, summed over all the stimuli. D_i reflects an individual's internal consistency in responding repeatedly to the same stimuli. The contribution that is made to this component by each stimulus is essentially the precision of the individual's score on each item, and when summed over the items is a measure of the *precision* of the individual's total score on the test.

The T_i component describes the variability of the individual's mean position within the group as the group passes from stimulus to stimulus. We call this score the individual's trait score.

Thus, we now have two concepts to represent the hypothetical behavior of an individual in response to repeated independent presentations of a set of items. We have the concept of a dispersion score which represents the precision of an individual's final total score on the test. And we have the concept of trait score which represents the stability of an individual's position within the group in passing from item to item.

Reliability and Homogeneity.—We shall now identify D_i and T_i with the concepts of reliability and homogeneity, respectively. We have here precise definitions of concepts from a psychological rationale such that the concepts may be manipulated mathematically and are susceptible to rigorous logic.

We shall use the terms D_i , dispersion score, precision, and reliability interchangeably; and the terms T_i , trait score, and homogeneity interchangeably. First, it is apparent from the mathematical definition of the concept of precision that it is a characteristic of an individual's behavior on the items comprising the test, and does not necessarily have the same value for every individual who takes a particular test. To put this in the more common terms of test theory, the reliability of a test or, as we define it, the precision of an individual's test score, may be different for every individual who takes the test. It is an approximation of unknown degree to assign the same coefficient

Acc No 2737

Date

to all individuals. This approximation, perhaps, would be reasonably close in the case of some mental tests, but in others the individual differences in D , might be considerable.

The relation between reliability and homogeneity is an interesting one. In principle we could construct a test which would have high precision, or reliability, and such that the items would have zero intercorrelations, or, for that matter, any values from plus one to minus one. Thus, if a man's score on one item was the number of children he has and on another item his cephalic index, and on a third item the number of clubs and societies he belongs to, his total score would have very high reliability. It does not necessarily follow, however, that the score means anything—that it represents a point on a continuum which is a psychological trait continuum. Obviously, then, the fact that one has high precision for a test score has no bearing on whether or not one is measuring some kind of meaningful psychological entity. If one takes a number of things which are qualitatively different and adds up the scores on these different things for each individual, then the total scores will be a set of numbers which may have the property of precision but will have no common quality.

Let us turn now to the trait score which we identify with homogeneity. This denotes the stability of an individual's position within a group. Such a measure would not be an exclusive property of an individual, as in the case of precision, but is a property of the group as a whole on the test, and hence T_i should be averaged over the individuals.

The significance of this concept lies in its indicating the degree to which the final total scores of individuals have some common quality or represent a psychological entity for the group. The expression for the trait score, T_i , averaged over individuals, is essentially equivalent to the notion of correlation between items, except that it is expressed in terms of variance rather than correlation or covariance.⁷

Thus, if we have a test consisting of a number of items, each

⁷ Another way of looking at D^2_i and T^2_i is by analogy with error variance and true variance in conventional test theory. The analogy between D^2_i and error variance is justified. But T^2_i is a variance generated by lack of homogeneity among the items. Hence, in the sense used here, the "true variance" would represent the degree to which the items failed to constitute an organized and integrated common trait.

from a different primary mental ability, we would expect the position of the individual within the group from item to item to be variable. This is on the premise that there are intra-individual differences in ability. On the other hand, if the test were a set of arithmetic items then the position of the individual within the group as it passed from item to item would probably be relatively stable and there would be a high degree of homogeneity. These two tests might well have equally high reliability but quite different homogeneities.

In principle, the two components D_i and T_i are independent and it is not difficult to imagine a test with perfect precision for all individuals, or perfect reliability, and with a degree of homogeneity anywhere from zero to perfect. On the other hand, in a probability sense, it would perhaps be much more difficult to construct a test with perfect homogeneity but with low precision. Such a relation is implicit in the reasoning behind the attempt to increase the reliability of a test by means of an item analysis against an internal criterion.

Indices.—We have reached a point now where we must consider again the distinction between the defined meaning of a concept and the index which presumably is a measure of the concept. What we have tried to do is to provide meaningful definitions of the concepts of precision and homogeneity but we have *not* provided an *index* for either one of these concepts. An index is simply a method of analyzing data to get certain information. Hence, in order to compute a meaningful index, the data must contain this information. Consider, for example, what is required of the data so that they will contain information about the precision of an individual's score. We can see that to get a measure of precision, that is, to compute an individual's dispersion score, requires repeated independent responses from him to the same item. The method of single stimuli conventionally used in mental testing does not provide such observations. Thus, it appears that with conventional testing methods an index of the reliability of a test score is indeterminate and there is no valid formula for reliability. On the other hand, the T_i component of an individual's total variance requires only one observation per individual per stimulus and, hence, data collected by the method of single stimuli

do contain information pertaining to the concept of homogeneity. But samples of size *one* are poor estimates of the mean of a distribution. Nevertheless, they can be used to get an estimate of the variance between distributions which is, however, contaminated by the variance within the distributions. The two components, D_i and \mathcal{T}_i , of the total variance cannot be separated in data collected by the method of single stimuli. In other areas, a method for collecting data like the method of paired comparisons or the method of triads does provide information pertaining to both components and it is possible in principle to measure them both.

Essentially, what we have done is to give the quantitative definition of concepts based on a psychological rationale precedence over the statistical procedure of computing an index and then arguing about what the index means. We have chosen to have meaningful concepts and to recognize that our measures of them are inadequate and approximate rather than to take the measures as experimental facts and try to give them psychological meaning with consequent ambiguity and controversy.

What is it, then, that we do get from our indices of reliability or homogeneity? It is apparent that we can have no clear index of either the precision of a test score or the homogeneity of a test from conventional testing methods. Every index designed to represent one or the other actually represents a joint effect. The various indices merely differ in the nature of their approximation, then, to V_i , the left hand side of equation (8), summed over all individuals.

Inasmuch as this V_i is also the variance of an individual's score just as one of its components, D_i , is, one might ask what the difference is between them. The difference is that D_i , the variability within an individual, is the degree of precision of a score on the *test*. V_i , the left hand side of the equation, is the precision of the individual's score on the *attribute*, the domain which the sample of items represents. Obviously, the homogeneity of the items in a test has nothing to do with the precision of a score on the test. But, obviously, this same score, when regarded as an estimate of the individual's score on the domain or attribute of which the items constitute a sample, is dependent upon the homogeneity of the domain. The greater

the homogeneity of the domain, the more alike will be the scores of an individual on successive samples of items from that domain.

IV. *Next Steps*

As we see some of the implications of this for the further development of test theory, there appear to be three general alternatives, the first of which has two sub-alternatives:

1. Continue with the method of single stimuli as a method of collecting data. Then we can do one of two things: (a) make the necessary assumptions to achieve an interval scale and hence have numbers to manipulate,⁸ or (b) drop the assumptions which lead to an interval scale and substitute Lazarsfeld's latent structure analysis (4). The first sub-alternative above is to continue in the conventional manner. This will permit easily accomplished empirical studies in which we could rarely have firm confidence and unambiguous interpretation. The second sub-alternative requires going in an entirely new direction. Lazarsfeld's latent structure analysis is a non-metric theory for the scaling of data collected by the method of single stimuli. Obviously, his theory could be taken over bodily by test theorists, although from a practical point of view there are still computational hurdles. Such difficulties, however, are mere mechanical limitations and are not defects of the theory.

2. A second general alternative is to discover or to develop a new method for collecting data which would enable us to put the items in rank order for each individual as to how well he passed them and how badly he failed them. If we could collect such data we would then have data which, with very simple assumptions, contain information about metric relations between stimuli and individuals (1).

3. A third alternative is to discover or to develop a new method for collecting data which would be equivalent to the method of paired comparisons. This would require repeated independent responses to each stimulus. Such data would contain information on the metric relations between stimuli and individuals, and, in addition, information on the two compo-

⁸ A better sub-alternative here is to experimentally validate the assumptions of an interval scale if this is possible.

nents of precision and homogeneity, making a precise distinction between them possible.

V. Summary

We have tried to show that the assumptions required for an interval scale and the identification of indices with concepts are serious obstacles to the further development of test theory. We have then developed a rational basis for defining the difficulty of a test item for an individual and, from this basis, developed mathematical expressions for the concepts of reliability and homogeneity. It was then made apparent that the measurement of reliability and homogeneity from the analysis of data collected by the method of single stimuli is not possible, as such data do not contain the necessary information. Several alternative directions for the further development of test theory are pointed out.

REFERENCES

1. Coombs, C. H. "Psychological Scaling Without a Unit of Measurement." *Psychological Review* .. (in press).
2. Coombs, C. H. "Some Hypotheses for the Analysis of Qualitative Variables." *Psychological Review*, LIV (1948), 167-74.
3. Stevens, S. S. "On the Theory of Scales of Measurement." *Science*, CIII (1946), 677-80.
4. Stouffer, S. A. *et al.* *Measurement and Prediction*. Princeton: Princeton University Press, 1949.
5. Thomas, L. G. "Mental Tests as Instruments of Science," *Psychological Monographs*, LIV (1942), No. 3.
6. Thorndike, R. L. "Logical Dilemmas in the Estimation of Reliability." *National Projects in Educational Measurement*. Series I. Reports of Committees and Conferences. XI (1947), 21-40.

PROBLEMS IN MEASURING THE EFFECTIVENESS OF PROFESSIONAL EDUCATION

DONALD K. BECKLEY

Simmons College

IN the process of completing a recent study of the effectiveness of one area of professional education, a number of problems arose that may well be of interest to others planning investigations of a similar nature. For this reason, this article has been prepared to describe some of these problems and the methods by which they were met. The study concerned was made to ascertain the effectiveness of college training for executives in retailing in terms of selected objectives determined to be desirable. To do this, the performance of retailing graduates in respect to these objectives was measured by means of an achievement examination and compared with the performance of other groups.

Selecting Groups for Comparison

A question arose at this point of what groups to use for purposes of comparison. It was recognized that, as in other areas of professional and vocational education, objectives thought to be desirable might very possibly be attained by means of work experience as well as through formal college training. A study of this nature could be helpful in identifying those objectives that could best be taught by means of formal college training and those for which work experience itself was best suited.

In appraising the effectiveness of formal training, it was thus necessary to take into consideration both formal training and work experience as factors to be measured in respect to achievement of the selected objectives. In order to have these two factors appear in all possible combinations, it was necessary to find subjects in each of these four groups: (1) no training, no work experience, (2) training, no work experience, (3) work

experience, no training, and (4) both training and work experience. Subject groups who met these requirements were obtained through the use of these categories, in which the distinguishing characteristics are the presence or absence of the two factors:

1. Incoming students at the Simmons College Prince School of Retailing who have neither studied retailing in formal courses nor had extensive work experience.
2. Students who have completed the course in retailing at the Prince School of Retailing, but have not yet had extensive work experience.
3. Employees in Boston stores who are in positions of the kind graduates soon will be taking, but who have had no formal retail training.
4. Store executives and junior executives who have had a specified amount of store experience and also are graduates of the Prince School of Retailing.

Because two programs of retail training are offered at Simmons College where this study was made, it seemed appropriate also to consider educational level as another factor. Hence, within each of the four groups were two sub-groups, one consisting of students who had completed a four-year undergraduate college liberal arts program, and the other including those who had spent only two years in liberal arts study before beginning their retail training.

The purpose of the study, then, was to determine the strength of these three factors: (1) formal retail training, (2) retail-work experience, and (3) under-graduate-college education in respect to achievement of selected retailing objectives. The hypothesis to be applied was that groups initially comparable in all respects but differing in their treatment should reflect differences in achievement that are the result of that particular treatment.

The nature of the experiment can best be indicated by arranging the data in the following design:

No experience

	No training		Training	
	2 yrs. coll.	4 yrs. coll.	2 yrs. coll.	4 yrs. coll.
Experience	$N = 36$	$N = 30$	$N = 29$	$N = 28$
	$N = 29$	$N = 32$	$N = 12$	$N = 10$

The basic test-score data are presented in Tables 1 and 2, in which the letters refer as follows: (a) no training, no work experience, (b) training, no work experience, (c) work experience, no training, and (d) both training and work experience. Numeral 1 refers to students with four years of college preparation, and numeral 2 refers to students with two years of college.

It was recognized that any statistical design selected could not be adequately precise when uncontrolled variables still remained. In this study, intelligence of the subject was meas-

TABLE 1
Means of Scores on Retailing Examination

Group	N	Total	I	Test Scores II	III	IV	V
a-1	30	41.47	9.13	8.07	9.07	8.77	3.80
b-1	28	59.57	12.72	11.54	14.39	13.50	7.41
c-1	32	50.01	11.75	9.34	11.41	10.41	7.91
d-1	10	60.10	12.70	11.30	14.60	12.70	8.80
a-2	36	36.41	10.01	7.86	7.67	8.39	5.06
b-2	29	56.17	13.13	9.54	14.17	11.34	7.82
c-2	29	44.31	11.03	7.58	10.14	8.62	6.38
d-2	12	50.25	11.66	8.33	12.33	11.33	5.58

TABLE 2
Standard Deviations of Scores on Retailing Examination

Group	N	I	II	Test Scores III	IV	V
a-1	30	2.11	2.02	2.02	3.07	2.21
b-1	28	1.43	1.68	1.97	2.23	1.86
c-1	32	2.09	1.73	2.81	3.09	2.95
d-1	10	1.62	2.05	2.16	2.61	1.54
a-2	36	1.62	2.46	2.26	2.24	2.09
b-2	29	1.18	2.20	1.69	2.19	1.53
c-2	29	1.71	2.40	2.90	1.47	2.59
d-2	12	2.09	4.13	1.95	1.57	2.69

ured by the *Wonderlic Personnel Test*, and sex differences were eliminated by having only women as the subjects. Recognizing that the age levels of the two sub-groups differ by several years by definition, calculation of critical ratios indicated that none of the differences between the means of the various groups were significant, thus minimizing age as a factor here.

Selecting Subjects for Administration

Some difficulties were encountered in obtaining an adequate sample of subjects in all of the groups. Categories 1 and 2

consisted of incoming and outgoing students, hence they were readily available to take the retailing examination. Through the cooperation of a large Boston department store, a comparable number of subjects in group 3 was made available. In the case of group 4, with both formal training and work experience, obtaining subjects was more difficult. In order to have work experience comparable in amount and degree to that of subjects in group 3, it was necessary to select graduates of the School who had been working for approximately one to two years. Because of the small number of graduates with this amount of experience, the total number of possible subjects was definitely limited. A practical difficulty faced here was that most of the 34 eligible subjects lived away from Boston, and, in fact, covered most parts of the United States. It was not practicable to talk with them in person, or to administer the examination personally, as was done with the other groups, and the only feasible method of reaching them was by mail. A letter was sent to each of these people requesting her assistance and enclosing the examination materials together with detailed directions as to the procedures to be followed. A follow-up card was sent to those who did not return the completed materials by the date suggested, and the final return consisted of 22 cases.

Because of the nature of the questions asked in the retailing examination, it seemed unlikely that more than a few of the 84 objective questions could be answered readily through the use of notes or texts. In view of the explanation that the average scores of each group rather than individual scores were of interest in the investigation, it further seemed unlikely that any of the subjects would have sought to use outside help in answering the questions. In the case of the *Wonderlic Personnel Test* there was the question of whether or not the subjects had adhered to the specified time limit. Each score was checked in terms of the subjects' previous academic performance as a student, and any earlier intelligence scores available. In the case of two subjects whose earlier record did not seem to justify the very high intelligence scores received, deductions were made arbitrarily to make their scores approximate the mean of the group excluding these two scores, where they would not

influence the group computations. In all other cases, the scores received appeared by inspection of the records available to be entirely probable, and hence were accepted as having been done under the conditions specified.

Checking the Reliability of the Examination

When the examination in retailing had been constructed, the question arose as to what measure to use in determining its reliability. This question might better have been considered before rather than after the examination was made. Because of the conditions under which this examination in retailing was developed and was to be administered, it was not feasible to measure reliability through the use either of a retest or of equivalent forms. Thus, it appeared that some use of the split-half technique or application of the Kuder-Richardson formulae was appropriate here. Originally the split-half technique was rejected because the examination had not been properly planned for the measurement of reliability, and there would have been an item discarded from each of several sub-groups when the odd and even items were matched. The Kuder-Richardson formula number 20, which gives an estimate of the reliability of a test when the numbers of items, the standard deviation, and the average variation of the items are known,¹ has been described as superior to coefficients obtained by the split-half method, because any error due to bias in splitting a test is eliminated.²

Because the examination in retailing was divided into five sets of items representing the five objectives being measured, it was desirable to estimate reliability coefficients for each objective separately. Similarly, the four groups to whom the examination was administered were different, and also were treated separately. Except for group 2, students who were tested at the time they were finishing their course in retailing and thus a highly homogeneous group in respect to test performance, all groups had reliability coefficients ranging be-

¹ See Kuder, G. F. and Richardson, M. W., "The Theory of the Estimation of Test Reliability," *Psychometrika*, II (1937), page 158.

² See Jackson, R. W. B. and Ferguson, G. A., *Studies on the Reliability of Tests*, Bulletin No. 12, Dept. of Educational Research, Toronto, University of Toronto, 1941

tween .514 and .900. When reliability was calculated by the split-half method in spite of the objection mentioned earlier, the coefficients for group 2 were shown to be higher than originally calculated, and within the range indicated for the other groups, thus leading to the conclusion that the examination was adequately reliable for group use.

Planning an Experimental Design

Perhaps the most important problem in undertaking a statistical study is the selection of an experimental design with a sufficiently high degree of precision to answer the questions desired. The problem here was to select a design to indicate whether or not differences in gains among groups of students were greater than would be expected from the operation of chance factors alone.

A technique often used in investigations such as this is the matching of pairs. It would have been possible to match pairs of cases within each pair of groups in this experiment, but the unequal number of cases would have proved to be a disadvantage in that many cases in the larger groups would be left over after pairs were matched.

One technique appropriate for use in this type of experiment is the analysis of variance. As described by Lindquist,³ the variance of a sample can be analyzed into two components: the within-groups variance and the between-groups variance. If the hypothesis of random sampling is correct, the two estimates of variance would normally differ only by chance. The *F* test, known also as the variance ratio, indicates at the desired level of significance whether or not the estimated variances are larger than chance. If so, there is reason to believe the hypothesis to be false.

In this experiment, however, it seemed especially desirable to ascertain the strength of the relationship among the factors. This measurement was not available through the use of analysis of variance, and the Peters' regression technique was used. The covariance technique could be used to account for the initial lack of equivalence of groups and also in estimating the rela-

³ Lindquist, E. F. *Statistical Analysis in Educational Research*. Boston: Houghton Mifflin Company, 1941. Page 76.

bility of differences between the adjusted final means. The Peters' technique seemed preferable, however, because it provided an index of the strength of the relationship comparable to a coefficient of correlation. This involved the matching of the experimental and control groups through use of a regression technique which does not require pair-by-pair matching. This treatment made it possible to know whether the three experimental groups did better on the achievement examination in retailing than would be expected in view of their intelligence test scores. The hypothesis tested here was that there were no real differences produced by the factors introduced, and that any differences in final mean scores, after allowances had been made for chance differences in initial mean scores, were due entirely to chance fluctuations in random sampling.⁴

This technique has been described by Peters⁵ as follows:

The method involves setting up a regression equation in rectilinear form based on the statistics of the control group, then predicting by it what should be the achievement scores of the members of the experimental group if they were just like the control group members; if, that is, the experimental factor produced no differential effect. We can, then, determine the differential effect for the experimental factor by the extent to which the average achievement of the experimental group exceeded or fell short of that predicted for it by the regression equation.

While similar in many respects to Fisher's covariance technique, the Peters' technique makes the regression equation from the statistics of the control group rather than from the experimental and control groups pooled, on the ground that a pooled estimate would be a meaningless hybrid if the two groups differed by reason of the experimental factor, as probably would be the case.⁶

The use of the regression technique is especially appropriate here, since it is recognized that there is a positive correlation between academic aptitude or intelligence, particularly verbal ability, and scores on the retailing examination. In this experi-

⁴ *Ibid.*, p. 181.

⁵ Peters, C. C. "A Method of Matching Groups for Experiment with no Loss of Population." *Journal of Educational Research*, XXXIV (1940), 70-74.

⁶ Peters, C. C. *et al.* "Research Methods and Designs." *Review of Educational Research*, XV (1945), 377-393.

ment, a regression equation was calculated from the scores obtained on an intelligence test and the retailing examination by the control group with neither retail training nor work experience. This equation was then used to predict the retailing examination score from the intelligence-test score for those in each of the three other groups, which were regarded as experimental groups. This predicted score was then compared with the actual score of each case in the experimental groups, and the significance of the difference between means of predicted and actual scores was tested for each objective in each sub-group separately.

A major problem in this connection concerned the standard error formula appropriate for use here. In this situation the different numbers of cases in the control and experimental groups do not affect the standard error formula, but the difference in the means of the matching scores of the control and experimental groups requires an adjustment for that difference. Thus, instead of the conventional formula for calculating the standard error of the difference between means, a special formula as stated by Peters and Van Voorhis⁷ must be used because the groups are not perfectly equated on the basis of the matching factors. The differences between the means attained in the various tests were then divided by the standard errors of the differences in order to determine the *t*-ratios.

The Peters' regression technique, described above, served to indicate clearly the level of significance of the mean differences in achievement scores when the groups were equated for intelligence, but they did not identify the relative strength of the factors being measured. The problem thus arose of how to measure the magnitude of the relationship between achievement in retailing and the several factors to be isolated: retail training, work experience, and college education. Some measure of correlation was needed here to indicate the strength of relationship between achievement and each of these factors with the other factors held constant.

The Kelley correlation ratio, ϵ , was found to be an appropriate statistical treatment for this purpose, particularly because it is not affected by disproportionate numbers of cases

⁷ Peters, C. C. and Van Voorhis, W. R. *Statistical Procedures and Their Mathematical Bases*. New York: McGraw-Hill Book Company, 1940.

in the various groups. As described by Peters and Van Voorhis,⁸ when corrected, ϵ has a standard meaning free from bias and independent of the size of the population of the sample and of the number of classes into which the sample is divided. It has been shown to have all the merits of analysis of variance, and, in addition, is interpreted positively rather than negatively, as in the case of the t - and F -scores involving the null hypothesis.

A problem, however, was how to set up the data in this study to make possible meaningful analysis. One plan used was to set up direct comparisons of various pairs of subject groups in order to isolate each of the three factors to be measured. For example, to isolate the factor of formal training, group 1 (no work, no training) was compared with group 2 (no work, training); and group 3 (work, no training) was compared with group 4 (work, training). By this kind of classification, direct comparisons were made between various pairs of groups, thus holding constant the factor present in or absent from both groups.

Although useful to some extent in measuring strength of relationship of the various factors, the ϵ treatment described above was not entirely satisfactory, and some further classification was sought whereby two of the three factors to be measured could be isolated simultaneously while the strength of the third factor was being measured. As described by Peters and Van Voorhis,⁹ there is a technique through which subjects can be sorted into classes on the basis of some known factor, and then subsorted into sub-classes. The variance of these sub-classes will be due to factors other than those which determine the class sorting. This treatment, which, in effect, is partial ϵ , was used in the study being described. Because two factors were to be held constant, it was necessary to sub-subsort the data. For example, to find the partial ϵ for education on achievement, with training and work experience held constant, the following classification was made:

Work Experience				No Work Experience			
Training		No Training		Training		No training	
2 yrs. c.	4 yrs. c.	2 yrs. c.	4 yrs. c.	2 yrs. c.	4 yrs. c.	2 yrs. c.	4 yrs. c.

⁸ *Ibid.*, p. 323.

⁹ *Ibid.*, p. 326.

Through the calculation of corrected r , it was possible to compare directly the strength of the three factors, and thus to have some statistical basis for noting the relative importance of these factors in respect to each of the objectives being measured.

The conclusions reached from this study were as follows:

1. The theory of retail training proposed that work experience alone can be more effective than formal training alone in teaching specific job techniques was not substantiated in respect to the objective: cultivation of skills in the use of retailing mathematics. Formal training alone was found to be approximately equal in effectiveness to work experience alone in this area.

2. The presumption that the combination of formal training and work experience together would prove more effective than either training or work experience alone was not consistently borne out, possibly because of limitations in the size of the sample studied. Although subjects in this group performed significantly better than the control group in the case of all but two sub-groups, these subjects did not consistently show significantly greater differences as compared with subjects with training or work experience alone. Many of the subjects tested had been working since graduation in personnel positions which did not directly involve customer contact or the use of merchandising mathematics, and the data suggest that as with training in other fields, people remember best those kinds of learning with which they are most directly interested or employed.

3. Of the five objectives measured, work experience was shown to be relatively the most effective in: (1) skill in the use of retailing mathematics, and (2) identification of retailing facts. Work experience was least effective in teaching the comprehension of the nature of distribution. As indicated above, work experience equalled formal training in effectiveness only in respect to skill in the use of retailing mathematics.

4. Subjects with four years of liberal-arts-college education were better prepared to be effective retail executives than those subjects with two years of liberal-arts-college work, except in the case of the objective: application of principles of retail management, where no significant relationship exists.

THE CONCEPT OF VALIDITY IN THE INTERPRETATION OF TEST SCORES

ANNE ANASTASI

Fordham University

If asked to define "validity," most psychologists would probably agree that validity is the closeness of agreement of a test with some independently observed criterion of the behavior under consideration. It is only as a measure of a specifically defined criterion that a test can be objectively validated at all. For example, unless we define "intelligence" as that combination of aptitudes required for successful school achievement, or for survival on a certain type of job, or in terms of some other observable criterion, we can never either prove or disprove that a particular test is a valid measure of "intelligence." The criterion may be expressed in very broad and general terms, such as "those behavior characteristics in which older children in our culture differ from younger children reared in the same culture," but, however expressed, it defines the functions measured by the particular test. To claim that a test measures anything over and above its criterion is pure speculation of the type that is not amenable to verification and hence falls outside the realm of experimental science.

To the question, "What does this test measure?", the only defensible answer can thus be that it measures a sample of behavior which in turn may be diagnostic of the criterion or criteria against which the particular test was validated. Nor is there any circularity implicit in such a definition of validity, since a psychological test is a device for determining within a relatively short period of time what could otherwise be discovered only by means of a prolonged follow-up. For example, with a psychological test we may be able to predict within a certain margin of error which applicants will succeed on a given job or which students will be able to complete a medical course satisfactorily. Logically, the same information

could have been obtained, even more precisely, by hiring all job applicants or admitting to medical school all students wishing to enroll, and observing the subsequent performance of each subject. The latter procedure is obviously so time-consuming and wasteful, however, as to be completely impracticable. Hence the tests make a real contribution in permitting predictions in advance of lengthy observations. Another advantage of standardized psychological tests is that they make possible a comparison of the individual's performance with that of other persons who have been observed in the same sample situation represented by each test. In other words, the tests provide norms for evaluating individual performance.

Prediction and comparison with norms represent valuable contributions which psychological tests can render to our knowledge of individual behavior, the practical benefits of these contributions having been widely demonstrated. It is of fundamental importance, however, to bear in mind that psychological tests do not provide a different *kind* of information from that obtained by any other observation of behavior. The use of such labels as "intelligence," "aptitude," "capacity," and "potentiality" has probably done much to make test users lose sight of the empirical validation of tests. A number of current disagreements regarding the interpretation of test results and the susceptibility of tested abilities to training may be traceable to a failure to take due cognizance of validation procedures. Many test users apparently give only preliminary and possibly perfunctory attention to validation data, in order to reassure themselves at the outset that the test is "satisfactory." Their interpretation of the scores obtained with such a test, however, often takes no account of the validation data and is expressed in terms which bear little or no relation to the criterion.

Perhaps one of the most common examples of such an inconsistent treatment of test validity is provided by what we may call the argument of "extenuating circumstances." Let us suppose that a child obtains an IQ of 58 on a verbal intelligence test, and that the examiner subsequently finds evidence of a fairly severe language handicap in this child owing to foreign parentage. It is a common practice to conclude in such a case that the obtained IQ is not "valid," on the grounds

that the verbal content of the test rendered it unsuitable for testing such an individual. At this point we may inquire, however, "On the basis of *what criterion* is this IQ invalid?" Certainly the obtained IQ may be a valid measure of the behavior defined by the criterion against which the particular test was validated. It is very likely that the same language handicap which interfered with performance on this test will interfere with the child's behavior in other linguistic situations of which this test is an adequate index. The correspondence with the criterion may thus be just as close for this child as for children without a language handicap. In school, for example, the language handicap would probably interfere with the child's acquisition of important skills and information. The resulting academic backwardness, together with the original language handicap itself, would, in turn, affect certain aspects of job performance and other areas of adult activities. Conversely, any remedial efforts designed to eliminate the language handicap would produce an improvement, not only in the tested IQ, but also in the broader area of behavior of which this test is a predictor.

It should be added parenthetically that language handicap has been chosen as an example only for purposes of discussion. A number of other "extenuating circumstances," such as visual or auditory defects, emotional and motivational factors, inadequate schooling, and the like, could have served equally well to illustrate the point. Similarly, the discussion has been limited to intelligence tests, since it is chiefly in connection with these tests that many confusions regarding validity have arisen. The entire discussion applies equally well, however, to all types of psychological tests.

Specifically, how does the case cited in our illustration, as well as others of its type, differ from those in which no question is raised regarding the "validity" of the test or its applicability to the particular individual? First, in the present case the examiner has direct and certain knowledge regarding at least one of the factors which *determine* the subject's subnormal performance, viz., language handicap. In other cases, the principal determining factor might be inferior schooling facilities, parental illiteracy, cerebral birth injuries, a defective thyroid,

or any of a large number of psychological or biological conditions. Yet it is doubtful whether the IQ would be considered "invalid" in all of these cases simply because it proved possible to point to a specific condition as the determining factor in the poor test performance. To be sure, in many cases of low IQ, the examiner has little or no knowledge about the circumstances or conditions which lead to the intellectual backwardness. But such ignorance is obviously no more conducive to "valid" testing. Quite apart from the question of validity, the examiner should, of course, make every effort to understand why the individual performs as he does on a test. The fullest possible knowledge of the individual's pre- and post-natal environment, structural deficiencies, and any other relevant conditions in his reactional biography is desirable for the most effective use of the test data. But to explain *why* an individual scores poorly on a test does not "explain away" the score. There are always reasons to account for an individual's performance on a test. Language handicap is just as real as any other reason.

A second distinguishing feature of our example is that such a language handicap is usually *remediable*. The individual need not be permanently backward in intellectual performance, but with special training he may in large measure compensate for past losses in intellectual progress. Susceptibility to treatment is, however, a matter of degree. Many of the conditions determining intellectual performance, whether structural or functional, are amenable to change under special treatment. Moreover, conditions for which no effective therapy is now known may yield to newly developed treatments in the future. The distinction in terms of remediability is thus rather tenuous. Nor does such a distinction have any direct bearing upon the validity of a measuring instrument. A thermometer may be a valid index of fever, despite the fact that the administration of medicine will cure the fever.

Thirdly, some may point out that language handicap is not *hereditary* and may maintain that for this reason its influence upon test performance ought to be "ruled out." Such an objection contains a tacit assumption that psychological tests are primarily concerned with those individual differences

in behavior which can be attributed to heredity. Since the number of hereditary conditions which have been clearly related to behavior differences are extremely few, such a policy, if followed consistently, would mean the virtual cessation of psychological testing. Moreover, the connection between hereditary mechanisms and behavior is so remote and indirect as to render the distinction between hereditary and environmental factors in behavior largely an academic one (cf, e.g., 2). Above all, it should be noted that no *criterion* against which any psychological test has been validated is itself traceable to purely hereditary factors. Hence no such test has been proved to be a valid measure of individual differences in hereditary characteristics.

A fourth point to be considered is that of *comparability*. It may be objected that the individual who is handicapped by language difficulties, sensory deficiencies, or similar "extenuating circumstances" is not comparable to the validation group on which the test norms were established. The requirement of comparability in the application of psychological tests needs further clarification. If individuals are entirely similar in all of the conditions (psychological, physiological, etc.) which influence the behavior measured by a particular test, individual differences will disappear, all subjects receiving the same score. Obviously no test is designed to measure behavior independently of the conditions which determine such behavior—that would be a logical absurdity as well as an empirical impossibility. When the conditions in which the individual differs from the standardization group affect the test and the criterion in an approximately equal manner and degree, the validity of the test for that individual will not be appreciably influenced by the lack of comparability of the individual to the standardization group.

This question of "comparability" pertains not so much to the measurement of behavior as to the analysis of the etiology of behavior differences. It is only when attributing the observed individual differences in test scores to a particular factor or class of factors that the investigator must make certain that other contributing factors have been reasonably constant. For example, if a few individuals in a group have a language

handicap while the rest do not, we could not ascribe individual differences in performance within this group to structural differences in the nervous system, or to any other factor whose contribution to behavior we may be investigating. The same limitation would apply, however, if educational opportunities, family traditions, incentives for intellectual activities, or any other factor were not held constant. The fact that the influence of language handicap, sensory deficiencies, and a few other conditions is more readily apparent does not place such conditions in a different category. The question of comparability applies equally to all conditions other than the one under investigation.

A fifth consideration pertains to the use of test scores in *prediction*. Could an IQ obtained by a child with a language handicap serve as a basis for predicting the subsequent behavior of the individual? As long as the language handicap remains, the test score can provide an accurate prognosis of the child's behavior in situations demanding the type of verbal responses sampled by the test. It is only in this sense that *any* psychological test makes predictions possible. Within a certain margin of error, behavior can be predicted *under existing conditions*. But if, for example, any detrimental conditions such as poor schooling, sensory deficiencies, or the like are corrected, then performance on *both* test and criterion will show improvement. In discussions of test *reliability*, various writers during the past twenty-five years have pointed out that a psychological test should be expected to reflect changes in behavior at different times and under different conditions.¹ For test scores to remain constant when conditions affecting the subject's behavior have altered would indicate a crude and relatively insensitive measuring instrument, rather than a highly "reliable" one. The same logic applies to validity. If the subjects' test scores remain unchanged despite the modification of conditions which affect criterion performance, the test cannot have high validity.

Closely related to the problem of prediction is the *scope or breadth of influence* of any given condition upon the individual's behavior. For example, the presence of a loud, irregular

¹ Cf., e.g., 1, 4, 5, 6, 9, 10, 11, 12, 15, 18, 19.

noise during the testing would probably affect the score on that test, without influencing the individual's behavior in other situations. A toothache or a severe cold on the day of the testing would be further illustrations of narrowly limited conditions. In the case of these conditions, the prognostic value of the test for the individual would indeed be reduced, in much the same manner that holding an ice cube in the mouth would invalidate an oral thermometer reading of bodily temperature. Conditions such as language handicap, however, affect the individual's behavior in a much broader area than that of the immediate test situation. They may thus influence both criterion and test score in a similar manner.

The import of the above analysis is that validity should be consistently interpreted with reference to the *specific criteria* against which the given test was validated. It also follows that validity is not a function of the test but of the use to which the test is put. A test may have high validity for one criterion and low or negligible validity for another. The attitude that a good test has "high validity" and a poor test has "low validity" is still too prevalent among test users. Tests cannot be validated in the abstract, nor is the usual concept of validity itself universally applicable to psychological testing. It is only when tests are employed for predictive or diagnostic purposes that the correlation with an external criterion is relevant at all. In many investigations concerned with fundamental behavior research, tests are employed merely as behavior samples obtained under standardized (i.e., uniform) conditions, without reference to the correlations of these samples with other, "every-day-life" behavior samples (i.e., practical criterion measures). When the maze-learning behavior of white rats is tested, for example, the maze is not first "validated" against the rats' success in finding food in a grocery basement, or their ability to avoid contact with prowling cats, or any other criteria of achievement in the rats' extra-laboratory or workaday world. The investigator may quite reasonably argue that for the study of the particular principles of behavior which he is investigating, maze-learning is as "good" a sample of behavior as cat-avoiding, and that he has no more reason for validating the former against the latter than vice versa.

Fundamentally, any validation procedure provides a measure of the relationship between two behavior samples. As Guilford has recently expressed it, "In a very general sense, a test is valid for anything with which it correlates" (7, p. 429). The process can be regarded as irreversible only when one of the behavior samples has greater importance than the other for a specific purpose.² In such a case, the more important behavior sample is designated the "criterion." No basic difference exists between "criteria" on the one hand and "tests" on the other. They are merely different samples of behavior whose interrelationships permit predictions from one to the other. We *could* predict intelligence test scores from school achievement, although the process would be needlessly time-consuming. In such a case, the intelligence test scores would constitute the criterion.

The criterion is not *intrinsically* superior in any sense. It is well known, for example, that many commonly used criteria, such as school grades or job advancement, may be influenced by many factors "extraneous" to the quality of the individual's performance. Yet, if it is our object to predict such criteria, with all their irrelevancies and shortcomings, then the correlation of a given test with such criteria *is* the validity of the test in that situation. To be sure, the immediate criterion against which a test is validated may itself have been chosen as a convenient index or predictor of a broader and less readily observable area of behavior. For example, a pilot aptitude test may be validated against performance in basic flight training, the latter being in turn regarded as an approximate index of achievement in more advanced training and even possibly of ultimate combat performance. Such "successive validation" would be quite consistent with the relativity of predictors and criteria. It might be noted parenthetically that it is only when criterion measures are themselves used as predictors of further behavior that one may legitimately speak of the reliability and validity of the criterion itself (cf. e.g., 8).

² To be sure, when the relationship between the two variables is curvilinear, prediction will not be equally accurate in both directions, since $r_{xy} \neq r_{yx}$. In such cases, however, there is no a priori reason to expect that the correlation will be any higher when predicting the "criterion" from the "test" than when predicting the "test" from the "criterion."

Validation against a "practical" criterion is essential for many uses to which tests are put. It should not be assumed, however, that only tests which have been validated against some criterion considered important within a particular cultural setting can be used in behavior research. In order to be able to generalize from any obtained test score, we need only to know the relationships between the tested behavior in question and other behavior samples, none of these behavior samples necessarily occupying the preeminent position of a criterion. Thus, if the investigator is interested in the possible use of maze-learning performance as a basis for predicting the rats' behavior in other learning situations, he will have to correlate the subjects' maze-learning scores with their scores in a variety of other learning tasks. If a common factor is identified through these different learning scores, the "factorial validity" (7) of any one of the tests in predicting that which is common to all of them can be determined. On the other hand, if no single learning factor is demonstrated, then the area within which predictions can be made must be accordingly narrowed to fit the confines of whatever common factor does become evident. Investigations conducted to date on human subjects, for example, have failed to indicate the presence of a common "learning factor" (20, 21), and animal studies have revealed even greater specificity (cf., e.g., 14, 16, 17). But such specificity, if further corroborated, is an empirically observed fact whose discovery is useful in its own right in advancing our knowledge of behavior; it should not be construed as a weakness of the tests.

Whether we are dealing with common factors and "factorial validity" or with "practical validity" in the prediction of everyday-life criteria, the question of validity concerns essentially the interrelationships of behavior samples. In the latter case, one sample is represented by the test and another, probably much more extensive sample, by the criterion. In the former case, the different tests which are correlated constitute the behavior samples. Nor should the terminology of factor analysis mislead us into the belief that anything external to the tested behavior has been identified. The discovery of a "factor" means simply that certain relationships exist between tested behavior samples.

The common misconception that the criterion is in some mysterious fashion more basic than the test probably results, in part, from the belief that tests measure hypothetical "underlying capacities" which are distinguishable from observed behavior. Discussions of psychological tests often become hopelessly entangled because of the implicit supposition that tests can be validated against such underlying capacities as criteria. Any operational analysis of actual validation procedures reveals the futility and absurdity of such an expectation.

In this connection we may consider a monograph by Thomas (13), which sounds a note of acute pessimism regarding the use of mental tests as "instruments of science." Through a careful and systematic logical analysis, the author demonstrates the fallacies inherent in any attempts to interpret psychological tests as measures of "innate abilities," hypostatized "fundamental human capacities," and the like. He clearly recognizes that "the methodology of mental testing provides no way of operationally defining an ability and a performance as distinct. . . entities" (13, p. 75). But, in his final conclusions, the author seems to exhibit the same confusions which he had previously sought to eliminate.³ For example, in the attempt to evaluate the scientific usefulness of psychological tests, he raises such questions as the following: "Do two identical scores mean that the same kind and amount of psychological processes were employed? Do they mean similar sociological backgrounds of experience? Do they mean a qualitatively similar adaptation to the immediate test environment? Do they mean that comparable amounts of psychic tension were built up or that similar amounts of nervous energy were expended?" (13, p. 77). By way of reply he adds: "The achievement of such scientific meanings as these from the current methodology of mental testing is probably too much to expect, for test results at present are notoriously ambiguous in what they signify about the socio-psychological ingredients of the recorded performances" (13, p. 77).

³ These confusions in the fundamental argument do not detract from the value of certain more specific points discussed in this monograph, such as the limitations of ordinal scales, and the concepts of difficulty value and homogeneity in test construction. But these problems have also been analyzed by other writers, in a somewhat more constructive manner (cf., e.g., 3, 10).

Two weaknesses are apparent in such an argument. First, the testing of behavior is being confused with an analysis of the factors which determine behavior. Secondly, despite his earlier advocacy of an operational definition of "ability," the author now appears to be chasing the will-o'-the-wisp of "psychological processes" which are distinct from performance. He seems thus to be demanding that in order to be proper instruments of science, psychological tests should measure functions which by definition fall outside the domain of scientific inquiry!

In summary, it is urged that test scores be operationally defined in terms of empirically demonstrated behavior relationships. If a test has been validated against a practical criterion such as school performance, the scores on such a test should be consistently defined and treated as predictors of school performance rather than as measures of hypostatized and unverifiable "abilities." It is further pointed out that conditions which affect test scores may also affect the criterion, since both test scores and criteria are essentially behavior samples. The extent or breadth of such influences is a matter for empirical determination, rather than for a priori assumption. Moreover, the validity of a psychological test should not be confused with an analysis of the factors which determine the behavior under consideration. Finally, it should be noted that the distinction between test and criterion is itself merely one of practical convenience. The scientific use of tests is not predicated upon the assumption that criteria are a separate class of phenomena against which all tests must first be validated. Essentially, generalization and prediction in psychology require knowledge of the interrelationships of behavior, regardless of the situation in which such behavior was observed.

REFERENCES

1. Anastasi, A. "The Influence of Practice upon Test Reliability." *Journal of Educational Psychology*, XXV (1934), 321-335.
2. Anastasi, A. and Foley, J. P., Jr. "A Proposed Reorientation in the Heredity-Environment Controversy." *Psychological Review*, LV (1948), 239-249.
3. Coombs, C. H. "Some Hypotheses for the Analysis of Qualitative Variables." *Psychological Review*, LV (1948), 167-174.
4. Cronbach, L. J. "Test 'Reliability': Its Meaning and Determination." *Psychometrika*, XII (1947), 1-16.

5. Dunlap, J. W. "Comparable Tests and Reliability." *Journal of Educational Psychology*, XXIV (1933), 442-453.
6. Goodenough, F. L. "A Critical Note on the Use of the Term 'Reliability' in Mental Measurement." *Journal of Educational Psychology*, XXVII (1936), 173-178.
7. Guilford, J. P. "New Standards for Test Evaluation." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, VI (1946), 427-438.
8. Jenkins, J. G. "Validity for What?" *Journal of Consulting Psychology*, X (1946), 93-98.
9. Kuhlmann, F. *Tests of Mental Development*. Minneapolis: Educational Test Bureau, 1939.
10. Loevinger, J. "A Systematic Approach to the Construction and Evaluation of Tests of Ability." *Psychological Monographs*, LXI (1947), No. 4.
11. Paulsen, C. B. "A Coefficient of Trait Variability." *Psychological Bulletin*, XXVIII (1931), 218-219.
12. Skaggs, F. B. "Some Critical Comments on Certain Prevailing Concepts Used in Mental Testing." *Journal of Applied Psychology*, XI (1927), 503-508.
13. Thomas, L. G. "Mental Tests as Instruments of Science." *Psychological Monographs*, LIV (1942), No. 3.
14. Thorndike, R. L. "Organization of Behavior in the Albino Rat." *Genetic Psychology Monograph*, XVII (1935), No. 1.
15. Thouless, R. H. "Test Unreliability and Functional Fluctuation." *British Journal of Psychology*, XXVI (1935-1936), 325-343.
16. Van Steenberg, N. J. F. "Factors in the Learning Behavior of the Albino Rat." *Psychometrika*, IV (1939), 179-200.
17. Vaughn, C. I. "Factors in Rat Learning: An Analysis of the Intercorrelations Between 34 Variables." *Psychological Monographs*, XIV (1937), No. 69.
18. Wherry, R. J. and Gaylord, R. H. "The Concept of Test and Item Reliability in Relation to Factor Pattern." *Psychometrika*, VIII (1943), 247-264.
19. Woodrow, H. "Quotidian Variability." *Psychological Review*, XXXIX (1932), 245-256.
20. Woodrow, H. "The Relation Between Abilities and Improvement with Practice." *Journal of Educational Psychology*, XXIX (1938), 215-230.
21. Woodrow, H. "Factors in Improvement with Practice." *Journal of Psychology*, VII (1939), 55-70.

THE LOGIC OF SCALE CONSTRUCTION¹

EDWARD A. SUCHMAN

Cornell University

Most of the classifications used in the course of our daily communication with one another are not defined with any great exactitude. For ordinary purposes of communication, it is usually not necessary to formulate a set of rules to distinguish between those things which belong to a certain class and those which do not. Agreement as to what constitutes membership in a class of objects is common enough to permit understanding without resort to explicit classification schemes. People can talk and write about "beautiful women," "successful men," "good books" or "prosperous nations" without bothering to state the rules for their classifications. These "loose" classifications constitute an important part of our communicatory system.

The Need for More Precise Classifications

To the scientist, however, who must work with these classifications, such loose usage often proves inadequate. Scientific communication demands a more rigorous statement of the bases for the classifications used. One of the tasks of the scientist becomes the translation of the loose descriptive terminology of ordinary social intercourse into the more precise classificatory systems of science. To the scientist the statements of Mr. Jones to Mr. Smith that, "Mr. Brown is a successful lawyer," or "Mr. Greene is an anti-Semitic person," or "The United States is a prosperous country," present problems in definition. What is meant by "a *successful* lawyer," or "an *anti-Semitic* person," or "a *prosperous* country"?

The need for such precise definition becomes apparent in ordinary communication when there is a disagreement between Mr. Jones and Mr. Smith. This disagreement illustrates the

¹ The author wishes to acknowledge the valuable contributions of Paul F. Lazarsfeld and Louis Guttman to the present formulation of the problem.

problem of communicatory classification which the scientist is attempting to solve. Two persons who disagree on how to classify a third person or object find themselves faced with the difficult problem of defining the bases for their classification. To reach an agreement they are forced to tighten the loose classificatory system which usually suffices when both are in agreement. So long as both individuals agree that "Mr. Brown is a successful lawyer," they will feel little need to define what they mean by "successful." However, when a disagreement occurs, they are forced to state more precisely what they mean by "successful" or "unsuccessful." This transition from a loose classification to a more rigorous classification constitutes one of the most important tasks of the social sciences. How can this transition be accomplished?

The Problem of Scale Construction

The efforts of social scientists to define the meaning of some attribute or variable in such a way as to permit the classification of persons or objects according to the degree to which that attribute is present or absent constitutes the problem of scale construction. As stated by Lundberg, there are two principal aspects to this problem, "(1) How shall we select the aspects or factors of a unit which we deem significant and which are therefore to be considered in our scale? (2) How shall we determine the relative weight to attach to each factor included?"¹ These problems of item selection and item weights occupy a central position in most current methods of scale construction.

However, we propose to show that in the case of a uni-dimensional scale these two problems are actually non-existent. The theory of "scalability" to be developed is based upon the fundamental concept that if an area is uni-dimensional, then (1) any series of items selected from that area is interchangeable with any other series of items, and (2) any set of weights given to a series of items will produce the same rank order of objects or individuals as any other set of weights. The problem of scale construction, therefore, takes the form of a test for uni-dimen-

¹ Lundberg, George. *Social Research*. New York: Longmans, Green & Co., 1942. p. 259.

sionality, rather than the arbitrary treatment of non-scalable data as if it were scalable.

First, we will deal with the problem of item selection and, second, with the problem of item weights.

"Non-itemized" versus "Itemized" Classifications

87 An attempt to clear up the disagreement between Mr. Jones and Mr. Smith discussed above may take two different lines of development: (1) The introduction of additional judgments from other persons, or (2) the listing of those items which serve to characterize the different classes. The first approach, that of "non-itemized" judgments or ratings, represents an attempt to reach an agreement based upon the opinions of other judges, without attempting to characterize or describe further the basis for the judges' ratings. The second approach, that of "itemized" classification, requires the listing of a characterizing aggregate of items which serves as the basis for the classification to be made.

Let us see how these two approaches would apply to the present problem. As an example of the first approach, Mr. Jones and Mr. Smith could attempt to settle their disagreement as to whether Mr. Brown is a "successful" lawyer by asking a group of other people to classify Mr. Brown as "successful" or "unsuccessful." The basis for agreement using this method might be the proportion of judges rating Mr. Brown as "successful" or "unsuccessful." This form of classification we shall call a "non-itemized" classification.

As a second approach, Mr. Jones and Mr. Smith could attempt to settle their disagreement by asking each other exactly what they mean by "successful" or "unsuccessful." They would probably reply by pointing out certain characteristics of Mr. Brown which to each of them signify the presence or absence of "success." The classification of "successful" is expanded by the introduction of such items as "He has money," or "People listen to what he has to say," or "He has written many books," and other classificatory items characteristic of "success." As more and more specific items are added to the general classification, the loose definition takes on a more pre-

cise meaning. Specific actions or characteristics of Mr. Brown are mentioned which afford the basis for the development of classificatory techniques by means of which "successful" people can be distinguished from "unsuccessful" people. The basis for agreement using this method might lie in the number of characteristics indicative of success which Mr. Brown possesses. This form of classification we shall call an "itemized" classification.

Thus, the need for a more precise classification, we have seen, can lead to the use of "non-itemized" judgments or to the use of "itemized" aggregates of characterizing attributes. Both methods are currently being used by social scientists in their attempts to classify data. Each method has its own particular set of problems. The use of "non-itemized" judgments presents a solution to the problem based upon ratings without any attempt to produce a definition of the variable. The use of "itemized" aggregates of attributes, on the other hand, attempts a solution to the problem based upon a meaningful definition of the variable. It is this latter method which will constitute the main focus of the present attempt to arrive at a logical basis for scale construction.

Let us look at the first question, "How shall we select the aspects or factors to be considered in our scale?"

The Concept of an "Itemized" Aggregate

An aggregate of items consists of a series of items which have been selected as characterizing some object or person. These characterizing items, as we shall see, form the basis for a rigorous system of measurement. The transition from a loose to a more precise classification, which is the task of the scientist, is accomplished through the organization of these characterizing items into coherent systems.

The number of characterizing items that exist for any single variable is unlimited. Furthermore, there appears to be little inherent reason why any one item is better than any other. "Success" may be defined in any number of different ways. Theoretically there are an infinite number of classificatory items which may be used to distinguish a "successful" from an "unsuccessful" person, no single one of which is inherently better

than any other. How can such an infinitely broad range of characterizing items be brought into the reach of the scientist who desires to study them?

This concept of a universe of items can be illustrated by examples from many different types of social phenomena. The construction of an index of purchasing power may include almost any sampling of characterizing items which come from the total universe of items characteristic of purchasing power. The classification of individuals according to social status may include a large group of characterizing items ranging from income to the number of books read. The judgment of individuals according to their ability to supervise men may include such diverse items as the amount of time spent talking to the men and the score received on an intelligence test. The intelligence test itself is composed of a wide range of items. The ranking of people according to their attitude toward some issue is based upon their responses to a series of attitude items. All of the above areas are characterized by the use of a wide range of items in an attempt to arrive at a more precise classification of these areas. Another way of stating this would be to say that an attempt is made to classify social phenomena by observing a number of items which come from a universe of items characteristic of these phenomena.

Sampling a Universe of Items

We now come to an important aspect of this concept of a universe of items—the sampling of items from this universe. Since an unlimited number of items can be used to characterize a single concept, any definite number of items that are used must be a sample from this unlimited universe. Any single item that is used in practice is but a sample of one from this universe, and is interchangeable with any other item from the universe that might have been used in its place. The items used in a scale of attitudes toward war, an intelligence test, a social-status scale, a rating sheet on efficiency of workers, a standard of living index, a personality inventory, or in any classification device in the social sciences are only a selection from an infinitely large number of similar items. Thus the practical problem of classification in the social sciences becomes one of study-

ing a universe on the basis of a sampling of items from that universe.

The concept of an aggregate of characterizing items, thus, conceives of a sample from an unlimited number of items which may be used to characterize any social phenomenon. The characterizing universe consists of all items which can be used to exemplify the social concept. The determination of whether or not an item belongs to a certain universe, however, remains a matter which must be decided upon by common agreement. A characterizing item belongs to a universe on the basis of some arbitrary decision as to its content. The universe itself is decided upon arbitrarily as the content of interest to the investigator. Some additional means, such as the consensus of judges, might be introduced to help the investigator, but the final decision of whether or not this item characterizes the universe or phenomenon of interest, must be a subjective one.

As will be discussed in the next section, a test of scalability can help one to eliminate certain obvious cases of misinterpretation of the meaning of an item. But such *ex post facto* rationalizations are to be rigorously avoided. If the decision is made that this particular series of items represents the universe of interest, then eliminating items must result in a redefinition of one's interests. Whether or not an item belongs to the universe must not be a decision based upon some "correlational" test—there must be an adequate "content" interpretation for both acceptance *and* rejection.

Our answer to the problem of which factors to consider in a scale, therefore, is that one must first define the universe in which one is interested. This definition of the universe is a subjective one and consists of the listing of characterizing aggregates of items. The actual series of items that one uses in practice can be conceived of as a sample of items from the unlimited number that exists in the universe of content. The problem now becomes one of determining how valid a representation of the total universe the selected sample is. The answer to this problem depends upon the determination of the dimensionality of the universe. Does the universe consist of a single dimension? To answer this question, we turn next to a consideration of "dimensionality."

The Concept of a "Uni-dimensional" Aggregate of Items³

Let us assume that we now have a tentative set of characterizing items to be used for the classification of some social phenomenon. What are the different patterns of inter-relationships which these items can assume and of what importance are these patterns to the problem of scale construction?

As an example of what might occur in the way of inter-relationships, let us start out with a simple case of three items only. Suppose, for example, in the previous problem of classifying individuals according to how successful they are as lawyers, we had decided to use the following three items:

1. Did he have an income of over \$25,000 a year?
2. Was he the author of any books on law?
3. Had he ever received any honors from the bar association?

Suppose further that each item had been answered either "yes" or "no."

Conceivably then we might have the following eight types occurring among the lawyers whom we are interested in classifying:

Type	Item 1 (Money)	Item 2 (Books)	Item 3 (Honors)
1	Yes	Yes	Yes
2	Yes	Yes	No
3	Yes	No	Yes
4	No	Yes	Yes
5	No	No	Yes
6	No	Yes	No
7	Yes	No	No
8	No	No	No

We are now faced with the problem of ordering the above eight types according to how successful each type is as a lawyer. Types 1 and 8 give us no trouble; type 1, possessing all three of the characterizing items of success, is most successful; and type 8, possessing none of the characterizing items, is least successful. However, we find that types 2, 3 and 4 each possess two of the characterizing items of success. How are we to rank these three types relative to each other? Should we give least weight to "honors," and rank type 2 above types 3 and 4, or should we

³This concept of a "uni-dimensional" universe has been derived from the theory of scaling developed by Louis Guttman. See "A Basis for the Scaling of Qualitative Data," *American Sociological Review*, IX (1944), 139-150.

give least weight to "books," and thus rank type 3 above types 2 and 4? The same problem of weighting applies to types 5, 6 and 7 each of which possesses one of the characterizing items. We are faced with the need to make some decision as to how much weight to assign to each of the characterizing items. We shall therefore call any aggregate of items with the above pattern of inter-relationship, aggregates which present a *problem of relative weights*. Rank order for such a pattern cannot be determined without assigning weights to the different items. Furthermore, depending upon the relative weights assigned, this rank order can vary with different sets of weights. This we recognize as the second problem of scale construction--how much weight to give to each item.

We now come to an important question, "Are there any aggregates of items which do not present a problem of relative weights?" It is to be expected that an affirmative answer to this question would depend upon our ability to find an aggregate of items which formed a rather special pattern of inter-relationships.

Let us illustrate one such pattern by means of the previous example. Suppose we found that the relationship between the three characterizing items was such that only four out of the eight possible types actually occurred. There would be (a) the type that possessed all three characteristics, (b) the type that possessed characteristics 2 and 3 only, (c) the type that possessed characteristic 3 only, and finally (d) the type that possessed none of the characteristics. In other words, only types 1, 4, 5 and 8, as listed above, would be found to occur in actuality.

Let us repeat this listing of types including only the above four types.

Type	(Money)	(Books)	(Honors)
1	Yes	Yes	Yes
4	No	Yes	Yes
5	No	No	Yes
8	No	No	No

Under what conditions could we expect the occurrence of only the above four types? The answer to this question is found

in the pattern of inter-relationship between the items. First, we find that the types can be ordered, depending upon the number of characteristics each type possesses. *No two types have the same number of characteristics.* Second, we find that the items or characteristics can be ordered, depending upon the number of types that possess that characteristic. *No two items are possessed by the same number of types.*

The result of this ordering process of characteristics and types produces a definite pattern of inter-relationship. This pattern can be easily recognized if we separate the possession of a characteristic from its absence, and then order both characteristics and types according to frequency of occurrence. The result of such an ordering process is a parallelogram.

This pattern could be represented as follows:

Type	Has Money	Wrote Books	Received Honors	Does Not Have Money	Did Not Write Books	Has Not Received Honors
1	X	X	X			
4		X	X	X		
5			X	X	X	
8				X	X	X

where an X represents the characteristics of each type. A *parallelogram* pattern such as the above offers no problem in weights. No matter what weights were given to each of the items, the rank order of Types 1, 4, 5, and 8 would be the same, because each type possesses *all* of the characteristics of the type below it, and one more in addition. The rationale for such a pattern will become clearer after the following discussion.

Another method of deriving this special pattern of relationship between characterizing items which do not present a problem of weights would be by means of cross-tabulation. What form must a cross-tabulation between two items take in order for a rank order based upon these items to be independent of any weights the items might be assigned? One form, of course, would occur if these two items were perfectly correlated. There would be only two types of individuals in such a case—those with both characteristics and those with neither characteristic.

This perfect correlation may be represented by a fourfold

table as follows:

		Item 1	
		+	-
Item 2	+	N	0
	-	0	N

where + indicates the presence of that characteristic and - indicates its absence. On the basis of this type of relationship between items, all of the individuals in the + + cell may be ranked above those in the - - cell. No matter what weights are given to the items, the rank order will remain the same.

A second possibility is that individuals may fall into three of the cells of such a fourfold table, as follows:

		Item 1	
		+	-
Item 2	+	N	N
	-	0	N

Here again there is no problem of relative weights to be assigned the two items. Those individuals in the + + cell would receive the highest rank order, those individuals in the - - cell would receive the lowest rank order, while the only other group in the + - cell would fall in between the highest and the lowest ranks. Again no matter what weights were given to the items, the rank order would remain the same.

Finally, a third possibility is that individuals would fall into all four cells of the table, as follows:

		Item 1	
		+	-
Item 2	+	N	N
	-	N	N

While there is no problem of ranking in relation to the + + cell and the - - cell, we find that how the individuals in the other two cells were ranked would be completely dependent upon the relative weights given to the two items. Here, then, we have the

problem of the relative importance of items or the problem of weighting in scale construction.

A cross-tabulation between any two dichotomous items in a series, therefore, must have the following characteristic in order for rank order to be independent of item weights; all cross-tabulations between the items should result in the absence of any cases in one of the cells which represents a "positive" answer on one item and a "negative" answer on another item. This zero-cell, furthermore, must occur in the column which contains the lowest positive frequency. For example, a cross-tabulation between items 2 and 3 of the previous example would have to look as follows:

		Item 2 (Books)	
		Yes	No
Item 3 (Honors)	Yes	N	N
	No	o	N

There should be no individuals who have written books, but who have not received any honors. Any characterizing item that is the property of a lower rank must also be the property of all higher ranks, while the lower rank must lack the distinguishing characterizing item of the upper rank. Thus, since "honors" is a characteristic of Type 5, it must also be a characteristic of Type 4 (a higher rank), but Type 5 in turn must lack the distinguishing characteristic of the higher rank, in this case "books."

The parallelogram pattern which permits the determination of a rank order without presenting the need for assigning arbitrary weights to the various items will be called a uni-dimensional pattern. Such a uni-dimensional pattern can be determined empirically, first by ordering items according to ascending order of positive frequencies, i.e., "money" is a characteristic of fewest lawyers, and is therefore placed before "books" which in turn precedes "honors," and then by ordering individuals according to the number of characterizing items they possess. If, as a result of this ordering of items and individuals, the aggregate of items with which one is dealing forms

a parallelogram pattern then we can proceed to classify individuals according to a rank order which is independent of any weights which the items might be given. Such a rank order has the property of permitting one to derive from the rank order the exact characteristics of the individuals in that rank—since there is only one possible combination of items for any single rank order. Furthermore, the rank order has the quality that any individuals in a higher rank possess all the characteristics of the individuals in a lower rank, and at least one more in addition. This property of reproducibility of characteristics from a knowledge of rank order can only be present where the aggregate of characterizing items does not present a problem of relative weighting. It permits a more clear-cut rationale for ranking individuals along a single continuum than is possible when the rank order must be based upon an arbitrary decision of how much weight to assign each item.⁴

The aggregate of items which permit such a rank order which is independent of item weights will be called a "scale" and the universe of which the items are a sample will be called a scalable universe. Since the universe is scalable, any selection of items from that universe would result in the same rank order of objects or persons as any other selection. A scale in the present usage is therefore an aggregate of items which are so inter-related as to offer no problem of relative weighting.⁵

A test of "single meaning"

To a limited extent, scale analysis can be used as a test of the "meaning" of items in an effort to eliminate items which do not belong to the scalable universe. However, there must be an adequate "content" reason in addition to the "correlational" analysis. In many cases, the correct decision would be to label one's universe of interest as multi-dimensional, and therefore

⁴ Simple techniques for testing a series of items for unidimensionality based upon the determination of whether or not a parallelogram pattern exists have been developed. See, for example, Guttman, L., "The Cornell Technique for Scale and Intensity Analysis," EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, VII (1947), 247-280.

⁵ It is important to remember that many universes will be found to present a problem of weighting constituent items and that much work remains to be done in solving the problem of classification for such areas. Scaling is not a solution to the problem of weighting, but rather a selection of areas which do not present a problem of weighting.

not scalable, rather than to attempt to tease out a scalable subgroup of items which no longer reflects the desired universe.

Let us illustrate this problem of "meaning" by means of an example. Suppose, in the previous example of the classification of lawyers according to "success," the item, "Does he have an income of over \$25,000 a year?" had been asked, instead, as, "Does he have an income of over \$25,000 a year which he has earned honestly?" Whereas an answer of "No" to the former wording of the question has a clear-cut single meaning, the answer of "No" to the latter wording may mean either that he does *not* have a high income or that he has a high income but, in the opinion of the respondent, he has not earned it honestly. The response to this latter question depends upon the aspect or element upon which the subject focuses. The question can have more than one interpretation. Such questions have been called "double-barrelled," and their use for classification purposes is limited by the fact that different subjects may be responding to different aspects of the question.

While the presence of double-meaning is relatively easy to determine in the case of a single question, there is another type of double-meaning which is not so easily detected. As was discussed in the first section, the study of social phenomena involves the sampling of items from a whole universe of items characteristic of those phenomena. This use of an aggregate of items permits the occurrence of a new type of double-meaning—different meanings for the different items in the aggregate. This problem is quite different from that of double-meaning in the single question, as can be illustrated by the following example.

Suppose, in the selection of items characterizing a successful lawyer, we had carefully avoided any single items with possible double-meanings. But we now add a fourth item, "Does he have children?" We now have the following list of questions:

1. Does he have an income of over \$25,000 a year?"
2. Has he written books?
3. Has he received any honors?
4. Does he have children?

Let us assume that there are no double-meanings in any single one of the above questions. However, a new problem arises. This problem may be stated as, "Do all of the above questions

deal with the same topic?" This problem is different from the previous one stated as, "Does this *question* call for a response dealing with a *single* topic?" The new problem is one of determining the single meaning of a *series* of questions, each of which has been judged individually to contain only a single meaning. In other words, we must now determine whether or not the single topic studied by each of the items is the same single topic for all of the items. The problem of meaning for a single question is, "Does the individual question produce a response to only a single topic?", while the problem of meaning for an aggregate of questions is, "Is the single topic studied by each of the questions the same for each question?"⁶

The proposed parallelogram test for uni-dimensionality would serve to indicate in a series of scalable items whether or not any of the items did not deal with the same dimension indicated by a large majority of the items. Such double-barrelled items as "Does he have an income of over \$25,000 a year which he has earned honestly?" or such extra-dimensional items as "Does he have children?" would not conform to the parallelogram pattern.

Summary

The problem of scale construction has often been stated as involving (1) the problem of item selection and (2) the problem of item weights. The present paper offers a logical system for scale construction which answers these two problems in terms of a test for uni-dimensionality. Any series of items used in a scale can be conceived of as a sample of items from an unlimited universe of items dealing with the variable being studied. If a test of the inter-relationships of these items shows them to conform to a defined parallelogram pattern, then the rank order of objects or individuals based upon these items will be independent of item weights. Furthermore, in such a case, the rank order will pertain to the entire universe of items and any selection of items from that universe will produce the same rank order as any other selection.

⁶ This question of meaning, of course, could be stated the same for both single items and aggregates of items as follows, "Is a single topic only being studied?" The present formulation, however, is important for an understanding of the methods used to answer this question for a series of items.

Thus, according to this approach, the problem of scale construction becomes a problem of testing a series of items for uni-dimensionality. If the items conform to the prescribed scale pattern, then the problems of item selection and item weights are non-existent. This approach therefore involves a test for scalability in the area of interest, rather than the construction of some arbitrary scoring scheme. In this sense, *scales can only be constructed for uni-dimensional variables*. If the underlying variable is shown to be uni-dimensional, the rank order of objects or persons is independent of item selection and item weighting. If the underlying variable is shown to be multi-dimensional, then a meaningful single rank order is impossible.

It is the task of the research worker, therefore, first, to define his area of interest by listing those items which characterize the universe in which he is interested, and, second, to test these items for uni-dimensionality. If the test shows that the universe is not uni-dimensional, then he cannot construct a meaningful scale by arbitrary decisions of item selection and item weighting. If the test shows that the universe is uni-dimensional, then the problems of item selection and item weights are non-existent.

•

VALIDITY, RELIABILITY, AND BALONEY¹

EDWARD E. CURETON

University of Tennessee

It is a generally accepted principle that if a test has demonstrated validity for some given purpose, considerations of reliability are secondary. The statistical literature also informs us that a validity coefficient cannot exceed the square root of the reliability coefficient of either the predictor or the criterion. This paper describes the construction and validation of a new test which seems to call in question these accepted principles. Since the technique of validation is the crucial point, I shall discuss the validation procedures before describing the test in detail.

Briefly, the test uses a new type of projective technique which appears to reveal controllable variations in psychokinetic force as applied in certain particular situations. In the present study the criterion is college scholarship, as given by the usual grade-point average. The subjects were 29 senior and graduate students in a course in Psychological Measurements. These students took Forms Q and R of the *Cooperative Vocabulary Test*, Form R being administered about two weeks after Form Q. The correlation between grade-point average and the combined score on both forms of this test was .23. The reliability of the test, estimated by the Spearman-Brown formula from the correlation between the two forms, was .90.

The experimental form of the new test, which I have termed the "B—Projective Psychokinesis Test," or Test B, was also applied to the group. This experimental form contained 85 items, and there was a reaction to every item for every student. The items called for unequivocal "plus" or "minus" reactions, but in advance of data there is no way to tell which reaction to a given item may be valid for any particular purpose. In this

¹ This paper was presented in Denver, Colorado, September 7, 1949, at a meeting sponsored jointly by the Division on Evaluation and Measurement of the American Psychological Association and the Psychometric Society.

respect Test B is much like many well-known interest and personality inventories. Since there were no intermediate reactions, all scoring was based on the "plus" reactions alone.

I first obtained the mean grade-point average of all the students whose reaction to each item was "plus." Instead of using the usual technique of biserial correlation, however, I used an item-validity index based on the significance of the difference between the mean grade-point average of the whole group, and the mean grade-point average of those who gave the "plus" reaction to any particular item. This is a straightforward case of sampling from a finite universe. The mean and standard deviation of the grade-point averages of the entire group of 29 are the known parameters. The null hypothesis to be tested is the hypothesis that the subgroup giving the "plus" reaction to any item is a random sample from this population. The mean number giving the "plus" reaction to any item was 14.6. I therefore computed the standard error of the mean for independent samples of 14.6 drawn from a universe of 29, with replacement. If the mean grade-point average of those giving the "plus" reaction to any particular item was more than one standard error *above* the mean of the whole 69, the item was retained with a scoring weight of *plus one*. If it was more than one standard error *below* this general mean, the item was retained with a scoring weight of *minus one*.

By this procedure, 9 positively weighted items and 15 negatively weighted items were obtained. A scoring key for all 24 selected items was prepared, and the "plus" reactions for the 29 students were scored with this key. The correlations between the 29 scores on the revised Test B and the grade-point averages was found to be .82. In comparison with the Vocabulary Test, which correlated only .23 with the same criterion, Test B appears to possess considerable promise as a predictor of college scholarship. However, the authors of many interest and personality tests, who have used essentially similar validation techniques, have warned us to interpret high validity coefficients with caution when they are derived from the same data used in making the item analysis.

The correlation between Test B and the Vocabulary Test was .31, which is .08 higher than the correlation between the

Vocabulary Test and the grade-point averages. On the other hand, the reliability of Test B, by the Kuder-Richardson Formula 20, was $-.06$. Hence it would appear that the accepted principles previously mentioned are called in question rather severely by the findings of this study. The difficulty may be explained, however, by a consideration of the structure of the B-Projective Psychokinesis Test.

The items of Test B consisted of 85 metal-rimmed labelling tags. Each tag bore an item number, from 1 to 85, on one side only. To derive a score for any given student, I first put the 85 tags in a cocktail shaker and shook them up thoroughly. Then I looked at the student's grade-point average. If it was B or above, I projected into the cocktail shaker a wish that the student should receive a high "plus" reaction score. If his grade-point average was below B, I projected a wish that he should receive a low score. Then I threw the tags on the table. To obtain the student's score, I counted as "plus" reactions all the tags which lit with the numbered side up. The derivation of the term "B-Projective Psychokinesis Test" should now be obvious.

The moral of this story, I think, is clear. When a validity coefficient is computed from the same data used in making an item analysis, this coefficient cannot be interpreted uncritically. And, contrary to many statements in the literature, it cannot be interpreted "with caution" either. There is one clear interpretation for all such validity coefficients. This interpretation is—

"Baloney!"

RESPONSE SETS: A NOTE ON CONSISTENCY IN TAKING EXTREME POSITIONS

EDWARD A. RUNDQUIST
Owens-Illinois Glass Company

"A RESPONSE set is . . . any tendency causing a person to give different responses to test items than he would when the same content was presented in different form." Thus Cronbach (1) defines response sets in a recent summary of the wide range of situations in which such sets have been found.

In personality testing, response sets can be deliberate attempts to deceive, reflections of basic drives or traits, reflections of a particular frame of reference, or a temporary set brought about by a particular way of interpreting the directions. On just what a response set reflects and how consistently it reflects it, will depend the importance that is attributed to it. If a response set is transient and dependent primarily on the given conditions of an immediate situation, interest will be confined to controlling its influence so it does not interfere with the interpretation of test results. If, however, it influences behavior in a variety of situations over a long period of time, it would be worthy of careful study as a means of personality measurement.

Among other response sets noted by Cronbach, is the tendency to take the extreme positions on scales of the *Like—Indifferent—Dislike* or the *Agree—Undecided—Disagree* type. This note reports on the consistency of this tendency in two situations, one immediately following the other. In the first, 111 factory girls, all doing the same work, describe themselves by indicating how well each of 200 descriptive words and phrases apply to them; in the second, how well they liked or disliked each of 100 activities.

As Cronbach notes, to measure the consistency of any response set, the situations involved must allow equal opportunity for it to be called out, i.e., the situations must be equally

indefinite or unstructured. Whether the personality and interest items with their respective directions provide this may be judged from the material appended to this note. To the writer there seems at least approximately equal opportunity for the set to take extremes to operate. The fact that the two forms were presented in immediate succession, the personality form first, would increase the likelihood for the same set to operate while taking both forms.

Substantial individual differences exist in the tendency to take the extreme position. The mean and sigma for the 200 personality items are 74.95 and 37.28; for the 100 interest items, 37.47 and 14.24. To obtain these scores, the number of A and E responses (see key at end of paper) for each series were summed.

The correlation between this tendency on the two series of items is .40. This is significantly different from zero. (Sigma of an r of zero with an N of 111 is .1.)

There is, then, a real tendency for those who take extreme positions in describing their traits to take extreme positions in describing their interest. On the basis of a consistency represented by a correlation of .4 in two similar and immediately successive situations, it is hard to believe this particular response set is reflecting anything basic about the individual. It seems rather that it is largely a function of the type of material, interpretation of directions, mood, or some other temporary condition. At least with a consistency of .4, we would not expect a measure to be very useful in predicting a criterion such as behavior on a job. With personality and interest items of the kind dealt with here, it would seem more profitable to eliminate the operation of this response set rather than to attempt to use it as a measure.

Directions for Personality Items

On the following pages are words and phrases used in describing people. You are to describe yourself by indicating how well each description applies to you. Use the following key:

- Key: A. Describes me *perfectly* or *almost perfectly*.
 B. Describes me *unusually* well.
 C. Describes me *fairly* well.
 D. Describes me *some* but not very well.
 E. Describes me *slightly* or *not at all*.

Indicate your answer by putting the letter that applies on the line in front of each description. Suppose the word is "helpful." If you feel it describes you fairly well, you would place a C on the line before it, thus:

C Helpful

If you feel that this word describes you some but not very well, you would place a *D* on the line thus:

D Helpful

If you feel that the word describes you perfectly or almost perfectly, you would put an *A* on the line, thus:

A Helpful

Look at each word or phrase and decide how well it describes you. Do not worry about being consistent but consider each description by itself. Do not skip any.

- | | |
|--------------------------------------|--------------------------|
| 1. Cheerful | 24. Have high ideals |
| 4. Know my own mind | 28. Stubborn |
| 8. Cooperative (like to help people) | 41. Like to be different |
| 13. Restless (never still a minute) | 44. Jealous |
| 14. Worry about the future | 46. Always on time |

Directions for Interest Items

On the following page is a list of activities. Indicate how much you *like* or *dislike* each one. Use this key in indicating how much you like or dislike it.

- Key: A. Like a great deal
 B. Like some
 C. Neither like nor dislike
 D. Dislike some
 E. Dislike a great deal

You may not have done all the activities listed. Further, some require training which you may not have had. For these indicate how much you think you would like them if you tried them and if you had the proper training. Answer every item. Give your first reactions. Work rapidly.

- | | |
|-----------------------------|------------------------------------|
| 3. Work around machinery | 19. Trying out new cooking recipes |
| 4. Arrange flowers | 28. Read a book |
| 8. Teach English | 30. Tidy up the house |
| 10. Soft and slow music | 38. Look up words in a dictionary |
| 16. Visit a canning factory | |
| 18. Go to parties often | |

REFERENCE

1. Cronbach, L. S. "Response Sets and Test Validity." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, VI (1946), 475-494.

THE INTERESTS OF ART STUDENTS

WALTER R. BORG

University of Texas

Introduction

THE aim of this study is to attempt to answer the following questions:

1. Does the *Kuder Preference Record* differentiate an art group from general-population samples with respect to art-interest scores?
2. Is success in art courses significantly related to areas of interests as measured by the *Kuder Preference Record*?
3. Do Kuder profiles of groups of students specializing in different areas of art differ significantly?

Preliminary Study

The *Strong Vocational Interest Blank* was used by the investigator in a preliminary study of 85 upper division art-college students and was not found to be useful in differentiating levels of artistic ability or revealing individual art interests, although, as a group, the art students studied were above the norms for non-art groups. Only 38 per cent of the art group studied in the preliminary investigation received "A" ratings in art interest although this group was made up entirely of advanced art students. Seventy-two per cent of Strong's criterion group, consisting of 124 painters, 79 commercial artists, 20 sculptors, and 9 cartoonists, made "A" art ratings on the Vocational Interest Blank.¹ The mean score for this group is given as 176.80 with a standard deviation of 88.08². The mean for the group tested in the preliminary study conducted by the author was 86 with a sigma of 100. Great differences between the general makeup of Strong's criterion group and the art-college students tested possibly account for the differences in score. For

¹ Strong, E. K. *Vocational Interests of Men and Women*. Stanford Univ.: Stanford University Press, 1943. Page 730.

² Taken from norms supplied with artist scoring key for Strong test.

example, the average age of the criterion group is given by Strong as 42.7 years, with average education 11.9 grade, indicating a much more mature and somewhat less-educated group than the advanced art students tested.³ Because of the above findings, it was decided to use the *Kuder Preference Record, Form BB*, in the present study.

Present Study

A total of 427 students at the California College of Arts and Crafts at Oakland, California, were used as subjects in this study. Of this group 299 were men (median CA 22-8), and 128 were women (median CA 19-7). Only students having completed nine or more semester units of art work at the school were studied.

Grade averages in art courses were used as the criterion for art-college success. Reliability of art grades was computed by comparing first-semester grade averages with subsequent grades of 92 students having completed more than 45 semester units of art work. A correlation of .84 was obtained, thus indicating that art-course grades in this college are reasonably reliable.

The *Kuder Preference Record* is scored for nine areas of interest. (1) Mechanical, (2) Computational, (3) Scientific (4) Persuasive, (5) Artistic, (6) Literary, (7) Musical, (8) Social Service, and (9) Clerical. As each response constitutes a choice of one area over two others, a picture of relative interest is given and not absolute interest, as is the case with the Strong test. The Kuder test has several advantages for research. Probably most important is the ease of scoring and the possibility of analyzing the scores. It is also comparatively easy to construct norms for selected groups when using the Kuder test and this was considered to be a useful undertaking as norms given by Kuder for art students and artists are not as complete as could be desired.

Results

Scores earned on area five of the *Kuder Preference Record* are intended to indicate interest in art. The mean scores of the 427 students in the art group used in this study, are closely in

³ Strong, *loc. cit.*

agreement with the twenty one cases given in the test Manual. Kuder gives a mean score in art interest of 85.7, which equals a percentile score of 96. Men in the art group studied had a mean score of 87.5⁴ (99th percentile), with the women's mean being 85.52 (96th percentile) or almost identical to the women artists studied by Kuder. The art group showed considerably less variability than Kuder's norm group, the sigmas being 9.20 and 12.82 respectively.⁴

The correlation between interest scores on the Kuder art scale and art-course grade point average was found to be only

TABLE 1
Summary of Scores of Various Groups of the Kuder Art Scale

Group	n	Mean	Sigma
Women Artists and Art Teachers	21	85.7	12.82
Total Art Group studied			
Men	293	87.5	9.20
Women	163	85.52	12.82
Upper 27% of Art Group	115	88.26	6.85
Lower 27% of Art Group	115	85.91	8.51
Fine Arts Group	87	86.83	12.19
Commercial Art Group	29	86.11	9.12
Art Teacher Group	113	86.47	9.67

.08, which is not statistically significant (t equals 1.66). Comparison of the upper and lower 27 per cent of the subjects with respect to grade-point average in art courses revealed a small and significant difference between means. The mean of the upper group was 88.26, the lower group 85.91 while the critical ratio of the difference was 2.32. The difference in scores between men and women was also slightly significant in favor of the men, the critical ratio also being 2.32.

In comparing three groups of students specializing in different areas of art at the California College of Arts and Crafts, no significant differences were found in their scores on the Kuder art-interest score. The means for the commercial art students, fine arts students, and art-teaching students were 87.11, 86.83, and 86.47 respectively. This places the means of all three groups between the 97th and 98th percentiles in art interest. Scores on the Kuder Art Scale for the various groups tested may be found in Table 1. It may be concluded that the Kuder Art Scale is

⁴ Kuder, G. F. *Revised Manual for the Kuder Preference Record*. Chicago: Science Research Associates, 1948. Page 12.

valuable in differentiating the art group from the general population. The low correlation with art grades indicate that it is not useful as an indicator of the degree of talent within an art group. This correlation would probably be much higher in a more heterogeneous group.

In addition to an analysis of the performance of the art group as a whole, it was decided to compare the performance of the three art-area groups of commercial art students, fine arts students, and art-teaching students in detail and construct profiles for them. It was considered most practical to first study these three area groups without regard for sex because of the small number of cases. Thus, the groups were first considered as a whole, and then the commercial art groups and the men's teaching group which contain sufficient cases were studied with respect to sex.

Table 2 gives scores of the three art-area groups on the nine Kuder Scales. It will be noted that there are no significant differences among the three groups in art interest, all scoring above the 95th percentile for both men's and women's norms. The commercial art group scored significantly higher than the other two groups in mechanical interest, it was superior to the teaching group in scientific and clerical interest, and was superior to the fine arts group in persuasive interest.

Table 3 shows a comparison between men's and women's raw scores and percentile scores in the commercial art group.⁵ Although considerable sex difference exists, it will be seen from comparing raw scores and percentiles that these differences are markedly less than those given in the norms. For this reason it is probable that, until more complete norms are published, a comparison of raw scores would be simpler and more valid than conversion to norm-group percentiles when dealing with art students. In examining the performance of the art-teaching group it will be seen that this group scored significantly above the other groups in social service interest. In spite of this difference, the average score of the teaching group in this area is only 70.52 which is below the 50th percentile on the test norms, indicating that consideration of scores in all areas, regardless of percentile rank may be more useful in some cases

⁵ Percentile Scores taken from profile sheet for the *Kuder Preference Record*.

TABLE 2
Mean Scores and Standard Deviations of Students in Commercial Art, Art Teaching and Fine Art on the Kuder Areas

Group	N	Mec	Com	Sci	Per	Art	Lit	Mus	Soc	Ctr
Commercial Art Group	265	M 69.76 ^{1*}	24.77	51.33 ²	68.05 ³	87.11	49.9	23.63	59.42	46.07 ³
		SD 17.11	9.41	12.71	16.00	9.12	12.77	8.99	15.58	13.16
Art Teaching Group	109	M 65.65	23.42	46.47	67.43 ³	86.47	53.15	25.57 ^{3*}	73.52 ³	41.21
		SD 16.4	9.26	13.4	16.21	9.07	12.7	8.61	17.32	13.59
Fine Art Group	57	M 61.05	25.57	49.32	61.10	86.83	58.93 ^{1*}	26.75 ¹	59.33	44.32
		SD 17.47	9.84	14.14	12.83	12.1	15.09	8.38	17.61	14.25

¹ Significant difference from Commercial Art Group mean.

² Significant difference from Art Teaching Group mean.

³ Significant difference from Fine Art Group mean.

* Significant at the 5% level only (others significant at the 1% level).

than restricting attention to extreme scores. The male art-teaching students were considered separately and their average scores did not differ markedly from the entire teaching group in any of the nine Kuder areas, thus giving some justification for using the same raw-score norms for both sexes until more complete norms are established by further research.

The fine arts group scored significantly above the other two groups in literary interest and also scored highest in music interest, being significantly above the commercial art group. Because of the small size of the group no sex differences were computed. Some data which may be of help in evaluating the performance of art with respect to raw scores on the *Kuder Preference Record* may be found in Tables 2 and 3.

TABLE 3
Comparison between Commercial Art Group Raw Scores and Percentile Scores on the Kuder Art Scale

Group	Mec	Com	Sci	Per	Art	Lit	Mus	Soc	Cle
Men's Raw Scores . . .	73.55	24.99	51.73	69.87	87.15	51.45	21.05	56.04	45.53
Women's Raw Scores . . .	59.95	23.59	48.52	62.83	87.22	47.68	19.42	68.73	48.97
Men's Percentiles . . .	38	16	23	46	99	63	73	14	32
Women's Percentiles . . .	71	23	37	54	98	42	42	22	22

Summary and Conclusions

With regard to the questions stated in the opening paragraph, the following conclusions may be drawn:

1. The group of art students in this study scored very high on the Kuder Art Scale, the men averaging 99th percentile and the women 98th percentile, thus differentiating them adequately from the general population.

2. The correlation between art-course success and art-interest scores is not significant for the group studied. The homogeneity of the art students with respect to level of art interest in part accounts for this low correlation.

3. A comparison of interest profiles for commercial art, art teaching, and fine arts students reveal that significant differences do exist. The commercial art group is significantly supe-

rior to both other groups in mechanical interest, exceeds the fine arts group in persuasive interest, and is significantly above the art-teaching group in scientific interest and clerical interest.

The art-teaching group is significantly superior to the commercial art group in music interest and exceeds the fine arts group in persuasive interest. The chief characteristic of the art-teaching group, however, is its social service interest which is significantly above that of the other art groups.

The fine arts group was high in literary interest and low in persuasive interest, being significantly different from the other groups in both. The fine arts group also exceeds the commercial art group in music interest.

All three groups score highest in art interest, but are very similar, all averaging between 95th and 98th percentiles according to the Kuder norms. These findings agree quite closely with interest clusters suggested by Kuder in the test Manual. Further study is necessary before the norms found in this investigation can be regarded with complete confidence.

•

A FACTORIAL INVESTIGATION OF FLEXIBILITY¹

ROBERT W. KLEEMEIER

and

FRANK J. DUDEK

Northwestern University

IN a previous investigation, performance on certain tests which were designed to measure flexibility seemed to be influenced by the ingestion of Benzedrine sulfate to a greater extent than was performance on "non-flexibility" tests (3). This evidence was not strong but was, none the less, provocative. The present study was designed to investigate more thoroughly the nature of flexibility by subjecting modifications of these tests to a more rigorous analysis. For the purpose of this study flexibility is defined simply as the ability (a) to shift from one task to another, or (b) to break through an established set in order to perform a task. We have preferred to use the term "flexibility" rather than the word "perseveration," which has frequently been used to describe the abilities measured by tests of the general kind used here, because the latter term so often has associated with it specific theoretical connotations, e.g., Spearman's mental inertia, Muller and Pilzecker's usage as a memory phenomenon, etc.

In an attempt to make the results as unambiguous as possible it was decided to investigate only one type of performance, viz., performance in which *S* would be required to shift tasks. Only simple tasks were used in the hope that factors would be more easily identified. Tests were designed to measure numerical, perceptual speed, and verbal factors. Within each area the attempt was to make some of the tests factorially pure. One test in each area, however, was designed to measure flexibility by requiring *S* to shift from one simple task to another. It was anticipated that factors associated with number, perceptual

¹ This study was aided by a grant from the Committee on Research of the Graduate School of Northwestern University.

speed, and verbal abilities could be isolated from this battery. The important consideration, however, was whether or not a factor which was common primarily to the tests requiring shifts of tasks would also emerge. If those tests which required shifts of tasks appeared on an independent axis, regardless of the type of ability represented, there would be evidence for a factor which might be called "flexibility" common to different types of tasks.

Description of Tests

Thirteen tests comprised the battery analyzed in this study. All tests were speed tests and, with the exception of the *Same-Opposite Test*, were administered in two parts so that estimates of test-retest reliabilities could be made. All tests were answered on separate *IBM* answer sheets. The various tests were:

Single Digit Numbers Tests (SDN).—Each of these tests consisted of 120 items administered in two parts of 60 items each. The time limit for each part was 90 seconds. S's task was to indicate whether answers of the problems as given were right or wrong. Sample items from each test are:

1. *Subtraction*

$$1. 8 - 3 = 6$$

$$2. 7 - 4 = 3$$

$$3. 3 - 2 = 2$$

2. *Addition*

$$1. 7 + 2 = 9$$

$$2. 5 + 6 = 12$$

$$3. 8 + 2 = 11$$

3. *Mixed*

$$1. 9 - 4 = 13$$

$$2. 8 + 4 = 12$$

$$3. 3 + 6 = 10$$

$$4. 3 - 2 = 1$$

$$5. 8 + 5 = 12$$

$$6. 7 - 1 = 6$$

Tests were administered in the following order: Subtraction (Part I); Addition (Part I); Mixed (Part I); Mixed (Part II); Addition (Part II); Subtraction (Part II).

Two Digit Numbers Tests (TDN).—These three tests were

the same as their counterparts in *SDN* tests, except that each of the numbers to be added or subtracted consisted of two digits, e.g., $11 + 36 = 47$. In no case were the sums greater than two digits, although remainders were either one- or two-digit numbers. Two and one-half minutes were allowed for work on each part of the test. The tests were given in the following order: Addition (I); Subtraction (I); Mixed (I); Mixed (II); Subtraction (II); Addition (II).

Same-Opposite Test (SO).—This test was comprised of 60 of the more difficult pairs of words drawn from various forms of the *Army Alpha Test 6* (7). *S* indicated whether the words had the same or opposite meanings. The time limit for the test was two and one-half minutes. Sample items are.

1. acme-climax
2. ligature-band
3. abstruse-recondite

Word Completion Tests (WC).—Each of these tests consisted of 60 items. Each test was administered in two parts of 30 items each with a time limit of 90 seconds for each part. *S*'s task was to select the one letter from among five alternatives which formed a word when used with a given stem of three letters. None of the stems were three-letter words in and of themselves. The tasks and sample items from each test are:

1. *Add final letter.* (In this test *S* had to select the letter which made a common four-letter word when added to the *end* of the stem.)

- | | | | | | | |
|----|-----|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| 1. | cen | d | c | a | t | r |
| 2. | lam | r | b | m | f | w |

2. *Add initial letter.* (In this test *S* had to select the letter which made a common four-letter word when added in front of the stem.)

- | | | | | | | |
|----|------|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| 1. | ouns | f | h | n | b | |
| 2. | alf | c | u | t | a | k |

3. *Mixed.* (In this test letters might go either in front or in back of the stem to form a word. No cues other than the stem itself were given as to whether the answer was an initial or final letter.)

- | | | | | | | |
|----|-----|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| 1. | pur | q | m | j | e | t |
| 2. | ite | a | h | j | b | y |

The tests were administered in the following order: Final letter (I); Initial letter (I); Mixed (I); Mixed (II); Initial letter (II); Final letter (II).

Perceptual Speed Tests (PS).—Each of these tests consisted of 60 items administered in two parts of 30 items each. The time limit for each part was two minutes. Each item consisted of a line of 30 capital block letters. S's task was to count the number of times some particular letter appeared. This letter appeared from one to five times in each line. The number of times it occurred was also the answer for the item which was entered directly on the IBM answer sheet. Sample items from the various tests are:

1. "N" test. (In this test the number of "N's" occurring in a line of "M's" was counted.)
 1. M M M M M M M M N M M M M M M M . . .
 2. M N M M M M M N M M M M M N M M . . .
2. "W" test. (In this test the number of "W's" occurring in a line of "M's" was counted.)
 1. M M M M M M M M W M M M M M M M . . .
 2. M W M M M M M M W M M W M M M M W M . . .
3. Mixed. (In this test each line consisted of "M's", "N's", and "W's". At the beginning of each line the letter to be counted was indicated in parentheses.)
 1. (W) M M M M N M M M W M M N W M N N M . . .
 2. (N) N M M M M W N M N W M M M M M M . . .
 3. (N) M M M M M N N W M M M N M M W M W . . .

The order in which the tests were administered was: "N's" (I); "W's" (I); Mixed (I); Mixed (II); "W's" (II); "N's" (II).

Population

The test battery was administered to 205 college students. These were tested in groups ranging in size from eight to 48 S's. These tests were given to 104 S's in the order in which they are described above, while 101 were given the tests in the reverse order. Since mean scores of groups showed no significant differences attributable to order, all data were combined into one group.

Results

Table 1 shows the zero-order intercorrelations of the tests in the battery. It will be noted that all correlations are positive,

TABLE I
Intercorrelations, Means, and S. D.'S of Flexibility Battery
(N = 205)

Test	1	2	3	4	5	6	7	8	9	10	11	12	13
1. SDN (sub)	836												
2. (add)	793	852											
3. (mix)	586	634	649										
4. TDN (add)	610	643	667	907									
5. (sub)	603	639	684	894	908								
6. (mix)	147	167	240	320	319	347							
7. S-O	346	398	425	563	551	550	409						
8. W-C	341	408	428	592	581	581	428	739					
9. (init)	307	330	352	465	461	481	329	698	663				
10. (mix)	376	393	411	356	344	359	072	280	222	183			
11. P-S	424	421	451	382	413	413	132	333	262	299	690		
12. (W)	469	503	525	488	483	519	137	353	355	340	705	806	
13. (mix)													
Mean	71.2	75.2	69.3	65.9	58.5	60.1	32.5	30.8	30.9	16.8	30.0	31.8	23.0
S.D.	15.0	15.9	15.0	15.2	15.6	15.5	9.1	7.8	7.0	6.4	5.9	6.9	4.8

ranging in size from .908 ($r_{1,6}$) to .072 ($r_{7,11}$). Raw-score means and *S.D.*'s are also presented in this table. Examination of these means shows that alternation or mixed tasks in the number tests appear to be of the same order of difficulty as the straight addition and subtraction tasks. This is true for both the Single Digit and the Two Digit tests. This result would not have been anticipated from results of the Benze-drine study. In that study mixed addition and subtraction problems were markedly more difficult than either the addition or the subtraction tests alone. Tests used in the earlier study were just like the Single Digit Numbers tests except that *S*

TABLE 2
Factor Loadings and Communalities of Variables in Flexibility Battery

Test	Centroid Loadings						Rotated Loadings					
	I	II	III	IV	h_c^2	h_e^2	I	II	III	IV	h_r^2	h_e^2
1.	.727	-.306	.342	-.226	.791	.80	.334	.149	.799	.156	.796	
2.	.777	-.291	.350	-.259	.878	.87	.337	.196	.836	.160	.876	
3.	.801	-.248	.297	-.213	.837	.84	.368	.236	.780	.192	.836	
4.	.847	.222	.272	.236	.896	.90	.287	.417	.404	.691	.897	
5.	.854	.205	.292	.238	.913	.91	.293	.403	.426	.697	.916	
6.	.864	.198	.264	.229	.908	.91	.314	.417	.421	.678	.909	
7.	.367	.309	-.094	-.086	.246	.30	.057	.464	.064	.143	.243	
8.	.700	.388	-.277	-.223	.767	.75	.253	.815	.152	.109	.763	
9.	.694	.427	-.203	-.204	.747	.74	.198	.803	.174	.169	.743	
10.	.604	.354	-.262	-.211	.603	.61	.214	.731	.120	.077	.600	
11.	.551	-.429	-.298	.218	.628	.64	.776	.029	.129	.081	.626	
12.	.636	-.423	-.376	.233	.779	.78	.863	.105	.116	.081	.776	
13.	.713	-.408	-.323	.247	.840	.83	.876	.132	.177	.150	.839	

wrote his answer beside the problem instead of indicating on a separate sheet whether or not the given answer was correct. Apparently it is just this difference that accounts for the divergence in results obtained here. In both the Word Completion tests and the Perceptual Speed tests the mixed sections yield significantly lower scores than do the other sections.

Table 2 contains the centroid and the rotated factor loadings. Four factors were extracted from the intercorrelation matrix. No significant residuals remained in the fourth-factor residuals. As a matter of fact, the fourth factor itself contributes little to any correlation. Rotations of these factors were made to satisfy, insofar as possible, criteria of simple structure and positive manifold. It is apparent that those tests which had been

designed to measure flexibility (tests 3, 6, 10, and 13) do not group themselves along an independent axis, but rather can be accounted for in terms of number, perceptual, and verbal factors depending upon the type of test. The factors are relatively easy to identify according to the nature of the task.

Factor I (Perceptual Speed-P).—The highest loadings in this factor are found in the *PS* tests (tests 11, 12, and 13). It is not surprising that *SDN* tests (tests 1, 2, and 3) show some saturation in this factor. The simplicity of the problems, for college groups at least, is such that perceptual speed might well influence the speed with which correct and incorrect answers are recognized.

Factor II (Verbal-V).—Tests 7, 8, 9, and 10 show the highest loadings in this factor. Test 7, Same-Opposite, shows less loading in this factor than might be expected, but, none the less, its relationship to the word completion tests is unmistakable. The *TDN* tests show minor loadings in this factor. Some correlation between verbal and numerical factors has often been observed. This relationship may depend on the complexity of the numerical task, since the *SDN* tests show practically no loading on this factor.

Factor III (Single Digit Number [SDN]).—Clearly, tests 1, 2, and 3 have the highest loading in this factor. Considering the apparent similarity between the *SDN* tests and the *TDN* tests one might have expected the latter to exhibit higher loadings in this factor. The *TDN* tests, however, came out on a factor of their own.

Factor IV (Two Digit Number [TDN]).—This factor shows the highest loadings in the *TDN* tests. Relatively little of the variance of other tests in the complete battery can be accounted for on the basis of this factor.

Table 3 shows the per cent of the total variance of each test attributable to each factor. Thus, 64 per cent of the variance of test 1 (*SDN*, subtraction) can be accounted for by the *SDN* factor, 11 per cent by the factor *P* and only 4 or 5 per cent by both factors *V* and *TDN* combined. The sixth column (h_i^2) shows the communalities computed from the rotated factor loadings for each test, or the per cent of variance in each test accounted for by the four common factors. The specificity

(Sp) of each test (column seven) is the difference between the reliability of the test (r_{II} , column eight) and h^2 . (Since h^2 cannot be greater than r_{II} , when this occurs it is apparently due to slight errors in estimating one or the other value.) It will be noted that Sp is close to zero for all tests with the exception of the Same-Opposite test. Since the Same-Opposite test was a speed test and since it was given in only one part, no reliability was computed. However, it was assumed that the reliability might be near .80. If this estimate is not seriously in error,

TABLE 3
Factor Variance Accounted for by Factors Isolated in Flexibility Battery

Test	I	II	III	IV	h^2	Sp	r_{II}	E_v
1.	1116	0222	6384	0243	80	00	80	20
2.	1136	0384	6989	0256	88	00	87	13
3.	1354	0557	6084	0369	84	07	91	09
4.	0824	1739	1632	4775	90	00	89	11
5.	0858	1624	1815	4858	92	00	92	08
6.	0986	1739	1772	4597	91	02	93	07
7.	0032	2153	0041	0204	24	56	80*	20*
8.	0640	6642	0231	0119	76	02	78	22
9.	0392	6448	0303	0286	74	00	73	27
10.	0458	5344	0144	0059	60	02	62	38
11.	6022	0008	0166	0066	63	16	79	21
12.	7448	0110	0135	0066	78	06	84	16
13.	7674	0174	0313	0225	84	00	76	24

* Estimated.

then the test shows a high degree of specificity. Apparently the verbal factor isolated here is not the only one necessary to explain scores on a rather difficult word meaning test.

Error variance (E_v) is shown in the last column of Table 3. It is the difference between 1.00 (assumed total variance of each test) and the reliability coefficient (non-error variance).

Discussion

The analysis of the data presented above suggests strongly that no factor of flexibility need be postulated to account for differences in performance on these simple alternation tasks. Since Spearman spoke so enthusiastically about the factor of perseveration, relatively little evidence has been put forth to substantiate the hypothesis (5). Most previous investigators have attempted to investigate the problem by using batteries

of tests which had been designed to measure anything that might possibly be considered under the term perseveration. Notcutt, for example, used a battery of 15 tests designed to measure sensory perseveration, motor perseveration (both creative effort and alternation type), and associative perseveration (4). Cattell likewise has used batteries of fairly complex tests; indeed, Notcutt borrowed extensively from Cattell's tests (1). It would seem that the aim of these investigators differed from ours. They were attempting to identify a general perseveration factor which would influence the total behavior of the person. Their postulate, based on Spearman's "Law of Inertia," led them to hope that this factor would pervade all sensory and motor activities. It should be found in learning and would be an important factor of temperament and as such should influence feeling, attitude, apperception, and even the "natural rhythm" of the individual. Experimental results do not support such a general factor (2).

In this study we have steered clear of perseveration conceived of as an all-pervasive factor. This concept has been avoided by the adoption of the term *flexibility*, and with the use of simple tests scored in a simple way. At this level of simplicity it is quite evident that a general factor of flexibility does not exist. Perhaps it may be demonstrated if a more complicated series of shifts between equally well-established habit patterns were to be required of *S*. Thus within a single factor area, say Number, *S* could be tested in addition, subtraction, multiplication, division, and various combinations of these functions. To this could be added various ways in which the problems could be presented. Flexibility would be demonstrated if the mixed tests should group themselves along an independent axis.

Notcutt presents results which appear at first glance to be at variance with our findings (2). He states that in his battery of tests the alternation tasks "reveal a genuine though small factor." His method of analysis involved the averaging of the intercorrelations of five alternation tasks. The average correlation thus obtained was 0.181 ± 0.030 . The tasks required in these tests were such things as writing H's then Π 's, and writing *ABCD*, then *abcd*. The method used in this analysis seems

somewhat tenuous for this result to be accepted with confidence. It would be interesting to determine what factors would emerge from an analysis of his intercorrelation matrix.

Scoring.—Earlier writers have been greatly concerned about the problem of scoring alternation tests. Since they were interested in the hindering or facilitating effects of the alternation task as compared to the homogeneous task, scores were sought to express this relationship. Cattell criticizes the use of the simple difference score ($X - Y$) by pointing out that this score will be highly correlated with speed (1). Thus if X = the score on the critical task and Y = the score on the homogeneous task, a slow worker would get a smaller flexibility score than a fast worker even though both experienced an equal amount of interference. Therefore, he used the ratio X/Y in scoring his tests. This scoring is satisfactory as long as one is working with tests requiring S to overcome habitual sets such as Cattell's triangle, reversed letter, and cancellation tests. Thus, on his reversed letter test Y would equal the score on writing the letters *opqrst* and X would equal the score on writing these letters in the reversed order *tsrqpo*. Walker, Staines, and Kenna, however, point out that this method cannot be used for alternation tests such as those in the present study (6). To do so one would let Y = the combined score on the homogeneous tasks, e.g., Addition plus Subtraction, and X = the score on the mixed task. The scoring formula X/Y under these conditions can give an accurate picture of an interference effect only if the speed of work on the two homogeneous tasks is equal. If one of the tasks is inherently more difficult than the other, this method of scoring will show an artifactual interference effect even though none may exist.² They suggest, therefore, that the

$$\text{Interference Score} = E/A$$

where E = expected score on the alternation task,
and A = actual score on the alternation task.

² As the authors point out, if S could do 60 addition problems in one minute and only 30 subtraction problems in one minute, his rate for addition would be one problem a second, and for subtraction one problem every two seconds. If S were doing both addition and subtraction (mixed) in one minute he should do 40 problems, and in two minutes 80 problems. On the other hand, if he spends one minute doing addition problems and one minute doing subtraction problems, he will finish a total of 90 problems. Here the scoring formula $X/Y = 80/90 = .89$. The interference shown is an artifact.

Thus if 60 seconds is devoted to each task,

$$E = \frac{60}{T_1 + T_2}$$

T_1 = the time to do each unit of the first task and is given by the formula

$$T_1 = \frac{60}{\text{score on the first activity}}$$

and $T_2 = \frac{60}{\text{score on the second activity}}.$

While the logic of these scoring methods is sound, it does not seem necessary to resort to them in correlational studies. In fact their use makes it very difficult to determine the relationship between the flexibility and the nonflexibility tasks. Thus, while the use of such scores makes it possible to determine whether or not interference exists, it is impossible to determine from them whether or not the interference is uniformly experienced by all *S*'s (high positive correlation) or is a factor unrelated to performance on the nonflexibility tasks. On the other hand, should the E/A ratio fail to give indication of interference, it still does not seem admissible to conclude that interference was not a factor. Thus, using this ratio on the obtained mean scores on the *SDN* tests (Table 1), an interference score (E/A) of .53 is obtained and on the *WC* test we get an interference score of .92. In both tests, of course, if the mixed tasks and the single tasks were both equally difficult, E/A should equal .50. At first glance, therefore, it would seem that flexibility could be a factor on the *WC* tests but not on the *SDN* tests. This is an erroneous impression, for in spite of the high average score on the *SDN* mixed test, it is necessary to know how the mixed test correlates with both the Addition and the Subtraction tests. If these correlations were appreciably lower than the correlations between the Addition and the Subtraction tests, it would indicate that some factor other than ability to add and subtract entered into the mixed task to lower the correlations. These correlations are, however, essentially equal. The same may be said for the *WC* tests. In this case since r_{8-9} , r_{8-10} , and r_{9-10} are all about equal in magnitude, indications are that a factor of flexibility need not be postulated to account for the relatively slow performance on the Mixed tests. Our factor analysis, of course, confirms this throughout the battery. Thus, it is felt that in factorial investigations

scoring formulae such as those described above are not only unnecessary but that actually these procedures may obscure data which can give valuable information.

Summary and Conclusion

The purpose of this study was to investigate the nature of flexibility by factorial methods. A battery of 13 tests was constructed. These tests were designed to measure numerical, perceptual speed, and verbal factors. Within each area the attempt was to make some of the tests univocal (factorially pure). One test of each type, however, was designed to measure flexibility by requiring *S* to shift from one simple task to another. The tests were designed for machine-scoring and were speed tests. *S*'s were 205 college students. Test scores were intercorrelated and the matrix of intercorrelations was factorially analyzed. Four factors were extracted. These were identified as: (*P*) Perception, (*V*) Verbal, (*SDN*) Single Digit Number, and (*TDN*) Two Digit Number. Those tests which required a shifting of tasks could be accounted for on the basis of the above four factors; consequently the postulated factor of flexibility common to the different types of tasks was not necessary to account for the obtained results.

In reference to scoring, the position is maintained that the various difference scores and ratio-scoring techniques used by other investigators are not necessary in factorial investigations of flexibility and indeed may obscure the essential relationship between the flexibility and nonflexibility performance.

REFERENCES

1. Cattell, R. B. "Temperament Tests. II. Tests." *British Journal of Psychology*, XXIV (1933), 20-49.
2. Cattell, R. B. *Description and Measurement of Personality*. Yonkers-on-Hudson: World Book Co., 1946.
3. Kleemeier, L. B. and Kleemeier, R. W. "Effects of Benzedrine Sulfate (Amphetamine) on Psychomotor Performance." *American Journal of Psychology*, LX (1947), 89-100.
4. Notcutt, B. "Perseveration and Fluency." *British Journal of Psychology*, XXXIII (1943), 200-208.
5. Spearman, C. *The Abilities of Man*. New York: Macmillan Co., 1927.
6. Walker, K. F., Staines, R. G. and Kenna, J. C. "P-Tests and the Concept of Mental Inertia." *Character and Personality*, XII (1943), 32-42.
7. Yerkes, R. M. (ed.) *Memoirs of the National Academy of Sciences*, XV (1921), 207-230.

THE STANDARDIZATION OF THE MOORE EYE-HAND COORDINATION AND COLOR MATCHING TEST¹

JOSEPH E. MOORE
Georgia Institute of Technology

THE *Moore Eye-Hand Coordination and Color-Matching Test* was originally developed to measure the speed of eye-hand coordination of small children, and it was found in subsequent studies to differentiate clearly the same factor in adults. It was thought that if a test of eye-hand coordination could be devised which would stimulate immediate interest, it would prove valuable in measuring certain differences in young children in whom this type of learning has not occurred to any great extent, or has not become highly specialized.

In order to devise a test which would appeal strongly to young children it was decided to utilize their interest in marbles. The first test that was constructed was a bulky affair and difficult to manipulate. By a process of trial and revision the instrument has been markedly changed and, it is to be hoped, improved. The pre-school and the adult tests are identical except in length. The pre-school, or short form, has been used to test both white and Negro children as young as two years of age. Motivation is rather easy, since children generally take a keen delight in picking up the marbles and putting them in the holes.

¹ This project was made possible in part through a grant-in-aid allocated by a Research Committee at the Georgia Institute of Technology from funds made available jointly by the Carnegie Foundation and Georgia Institute of Technology. The author, however, and not Georgia Institute of Technology, is solely responsible for statements made in this report.

The writer wishes to acknowledge the assistance and cooperation of the following individuals: President Robert P. Daniel and Mr. William N. Smith, Personnel Counselor, Shaw University; Prof. Dorinda Duncan, Tuskegee Institute, Dr. Susan Gray, George Peabody College; Dr. Sidney Q. Janus and Prof. Albert S. Glickman, Georgia Institute of Technology; Prof. Herman Long, Fisk University; Dr. C. W. Thomasson, Drexel Institute; Dr. R. R. Ullman, Wittenberg College, and Prof. Joseph L. Whiting, Atlanta University.

The following picture shows the adult or long form of the test.

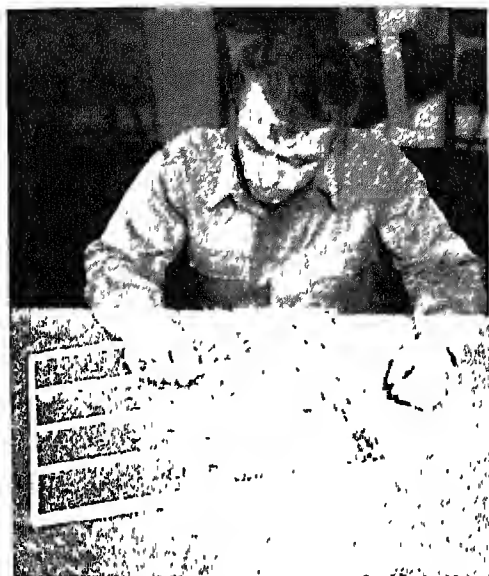


FIG. 1. Eye-Hand Coordination and Color Matching Test

It will be seen that the test is a rectangular board $16\frac{1}{2} \times 19\frac{1}{2}$ inches. The thickness of the test is slightly over $\frac{3}{4}$ inch. There are four rows of one-half inch holes. Each row contains eight holes spaced $1\frac{1}{2}$ inches apart. There are four starting boxes or slots holding the marbles, one at the end of each row of holes. Each box holds eight marbles for the Speed Test and twelve marbles for the color-matching part of the test.

The Color-Matching Test operates in the following way: Under each hole there is a colored piece of paper covered by transparent tape. The colors, in order, are red, green, blue, and yellow for the first row of eight holes. The second row is green, blue, yellow, red, the color sequence being different for each row.

The Pre-School Test is actually half as long (16 marbles in each trial are used instead of 32) as the adult form. The child is seated comfortably and is told to watch as the examiner shows him how to play the game. The examiner then takes one marble at a time and puts it in the hole so as to give the impression that it is fun to play the "game" fast. The child is then permitted to take a practice trial on the first eight marbles. The

test score is the total number of seconds it takes a child to do the 16-hole test three times, placing the marbles in consecutive order. A Pre-School Test can be made by covering one-half of the Adult Test with a piece of cardboard.

The norms for the Pre-School Test were based on the scores of children from nursery schools, kindergartens, and lower-elementary schools from the states of Tennessee and Kentucky.

From Table 1 it is seen that the average time for each age group becomes progressively faster. Comparison of the means with medians shows that every group except one is negatively skewed.² The range in scores indicates the extremes that are found in the reaction time of children within the age range studied. The standard deviation tends to become progressively

TABLE 1
Speed Measured in Seconds, of Eye-Hand Coordination of 431 Children on the Pre-School Form

	Age in Months									
	24-29	30-35	36-41	42-47	48-53	54-59	60-65	66-71	72-77	78-83
	Number of Children									
	10	25	45	56	47	54	49	38	78	29
Range of Speed	141-448	122-372	88-275	90-210	81-205	76-222	75-191	66-146	60-110	65-95
Median	215	177.5	142	135	120.9	106.3	100	88	83.4	79.1
Mean	225	177.5	156.7	137.6	123.3	112	108	93.5	84.6	80.9
Standard Deviation	94.7	58.2	48.7	26.7	29.4	27.0	26.2	22.5	10.7	7.6

smaller for each succeeding higher age group represented in the sample.

The Long Form or Test for Adults

The long form of the test requires placing 32 marbles, one at a time, in consecutive order in the holes. The test is taken in a seated position and has been standardized at typing-table height, or approximately 26 inches. The subject first has a practice trial of a row of eight marbles. The individual's score is the total number of seconds it takes him to complete three runs of 32 marbles each.

The long form of the test has been employed to measure the speed of eye-hand coordination of children in both elementary and high schools. The data on the performance of individuals

² Scores represent the number of seconds necessary to do the test. The fewer the seconds the faster the performance. The distribution therefore represents slow scores at the left and fast scores at the right.

between the ages of six years and sixteen years are presented in Table 2.

Each age group represented in the above sample (Table 2) completes the test progressively faster than the next younger age group. If the small sub-group samples are representative of the corresponding larger populations it would seem that the smallest changes in speed and precision occur in the younger age groups, ages six through ten, and the greatest between the ages of eleven to sixteen. It will be noted that four of the age groups are positively skewed, the median being larger than the mean. The age groups which are positively skewed are six, seven, thirteen, and sixteen. The small sampling could account for a part or all of the skewness.

TABLE 2*
Speed, in Seconds, of Eye-Hand Coordination for 602 Subjects Aged Six Through Sixteen Years

	Age										
	6	7	8	9	10	11	12	13	14	15	16
	Number of Subjects										
	33	58	66	42	18	38	45	43	28		207
Range	121-200	110-195	105-175	93-160	101-142	100-135	81-150	84-130	83-141	2 133	75-125
Median	166.3	145.0	133.0	130.0	120.0	116.5	115.9	110.8	107.3	103.8	95.4
Mean	162.6	144.5	135.0	132.9	123.3	118.1	116.1	109.7	108.9	104.5	93.3
S.D.	19.7	15.5	15.2	14.9	12.1	7.7	12.1	9.2	12.9	11.7	11.0

* Detailed norms are given in the Manuals.

Data available on the long form of the test indicate that it can be considered reasonably well standardized, at least on Southern men and women. The data from two Northern schools, Drexel Institute and Wittenberg College, are so similar that it does not appear that any great divergence of central tendencies and variability are to be expected in other areas. Further studies are encouraged, however, to prove the accuracy of this assumption.

The data that have been accumulated on the long form of the *Moore Eye-Hand Coordination and Color-Matching Test* are presented in detail for adult subjects in Table 3. Separate norms have been presented for white and Negro subjects. The justification for the separate norms for whites and Negroes was the fact that the difference between the average time of the two groups favored the whites on both the speed and the color-

matching tests. The difference in performance of the whites was statistically significant at the 1 per cent level for all comparison except that between white women and Negro women on speed, in which instance the difference was not statistically significant.

Table 3 reveals that on the speed of eye-hand reaction, women are faster than men. White men did the test more rapidly than Negro men and white women did the test more rapidly than Negro women. These differences favoring the

TABLE 3
Norms for Adults on the Long Form of the Moore Speed of Eye-Hand Coordination Test

	College	White Men Non-College	Busi. & Ind	Negro Men College	Non-College
Number of Subjects	776	2,707	1,222	451	108
Range	73-123	70-180	75-175	80-150	85-180
Median	96.18	104.30	104.04	100.92	109.0
Mean	96.72	106.00	103.26	99.20	111.0
S.D.	8.85	12.00	10.48	9.17	14.5

	White Women College	White Women Busi. & Ind.	Negro Women College
Number of Subjects	324	348	280
Range	74-144	74-131	64-136
Median	94.27	98.5	95.61
Mean	95.59	99.0	96.25
S D	9.55	9.25	10.15

white men are statistically significant. The difference between the mean of college women favored the faster performance of the white women but the difference is not statistically significant. Negro women performed the test somewhat more rapidly than did men in the college groups. The greatest differences in speed of eye-hand coordination are between college and non-college groups rather than between racial groups.

The non-college white males were men who came through the Georgia Tech Guidance Center and were being considered for work calling for some type of manipulative skill. The scores of these men were negatively skewed. In short, these men did fairly well in performance calling for quick and accurate manipulation insofar as such factors were measured by the Moore test. It will be seen that Negro college men also worked much faster than the non-college Negro group.

The Color-Matching Test

The Color-Matching Test has been developed during the last eight years at the request of certain industrial firms. The test requires that the individual match a marble of a specific color with a hole of the same color. The colors, as were mentioned previously, are arranged in an irregular order. The four colors used are red, green, blue, and yellow. The score on the color-matching part of the test is the number of seconds required to complete the test; that is, to match the 32 marbles with the 32 colored holes three times. If a mistake is made, such as placing a red marble in a yellow hole, *one* second is

TABLE 4
Norms for Adults on Speed of Color-Matching

	College	White Men Non-College	Ind. & Inv.	Negro Men College	Non-College
Number of Subjects	368	1,181	701	451	81
Range	87-170	90-349	90-304	91-258	100-280
Median	114.50	132.9	132.02	123.72	152.0
Mean	115.54	137.1	133.50	125.01	158.0
S. D.	12.20	24.8	18.42	15.35	36.0
		White Women College		Negro Women College	
Number of Subjects		322		280	
Range		80-214		84-190	
Median		110.00		118.20	
Mean		110.02		121.33	
S. D.		11.70		16.20	

added to his score for each such error. The subject is not permitted to arrange the marbles in a definite order previous to the starting signal.

Table 4 presents the data on the color-matching test with separate norms for white and Negro groups. The white college groups did the test more rapidly than the Negro college groups. The differences between the respective means were statistically significant at the 1 per cent level.

As would be expected, color-matching takes longer than the simple Speed Test. It takes an individual between five and ten seconds longer per trial to match the colors, or from fifteen to thirty seconds longer for the three trials. A comparison of Tables 3 and 4 reveals that the differences among the various groups are more pronounced on the Color-Matching Test than

on the simple Speed Test; especially is this the case in the comparison of college and non-college white men.

Validity

The validity of the *Moore Eye-Hand Coordination Test* was investigated in a number of ways. Age differentiation was one criterion. The test differentiated between the various age groups from 24 months to sixteen years and older. As each group of subjects took the test those who were older tended to make faster scores. After the sixteenth year age did not seem to have any appreciable relation to speed on the groups included in this study.

In the business and industrial field two studies are available on validity. In one study ten ice cream sandwich makers took the Speed Test and the scores were correlated with the number of dozens of sandwiches each turned out in a specified time. The coefficient of correlation was .52. The second study dealt with 23 loom operators or weavers. The group was divided into those above and below average on the speed of color-matching and above and below \$1.18 in hourly earning rate. A tetrachoric correlation of .86 was obtained.

The correlations of the Moore tests with other dexterity tests were also used as indirect evidences of validity. The speed of eye-hand coordination correlated .51 with the *Pennsylvania Bi-Manual Work Sample* (assembly) on 317 adult male subjects. On the *Minnesota Rate of Manipulation Test* the coefficient for placing was .67 for 157 subjects, and for turning it was .45 for 191 cases. The color-matching part of the Moore Test gave the following correlation coefficients with other tests: *O'Connor Tweezer*, .54 for 133 subjects; *Minnesota Rate of Manipulation* (placing), .53 for 103 men. The *Pennsylvania Bi-Manual Work Sample* (assembly) correlated .51 for 237 individuals and for disassembly, .50.

Reliability

The reliability of each of the three trials of the test was also studied on 441 men. The scores for trial one (32 marbles) were correlated with the scores for the second trial (32 marbles), and a coefficient of .83 was found. Scores for trial two were correla-

ted against scores for trial three and a coefficient of .77 was obtained. Scores for trial one were then correlated with scores for trial three and a coefficient of .82 was revealed. From these data it would appear that the instrument is doing a fairly consistent job of testing, even if the obtained speed on the first 32 marbles is taken as a criterion of actual speed.

The reliability of the color-matching part of the test was computed on the scores of 83 college juniors and seniors by the test-retest method. Total scores (96 marbles) obtained one week apart were found to give a coefficient of correlation of .82. When this is corrected for the restricted range, the coefficient becomes .955.

The *Moore Speed of Eye-Hand Coordination and Color-Matching Test* yielded a correlation coefficient of .67 on a group of 364 adults. It would appear that the speed element is playing a major part in both tests.

The reliability of the Pre-School Form was checked on 81 children drawn from the pre-school group and the first two grades of elementary school. The test-retest method after a period of one week gave a coefficient of reliability of .95.

Summary

1. The pre-school Form of the *Moore Speed of Eye-Hand Coordination Test* differentiates the performance of children between ages of 24 and 72 months. The reliability of the Pre-School Test for 81 children by the test-retest method after one week was .95.

2. The long form of the test is able to differentiate between each age group for ages six through fifteen years. After the sixteenth year speed does not appear to be very closely related to age for the groups included in this study.

3. The validity of the Speed Test has been investigated by using such criteria as age differentiation, correlation of the speed and production of a group of ice cream sandwich makers (.52), and correlation with the *Minnesota Rate of Manipulation* (placing) for 157 subjects (.67). On the Color-Matching Test a group of 23 weavers was divided into above- and below-average groups on the color-matching test scores and above- and below-average groups on hourly earning rate, and yielded a tetrachoric correlation coefficient of .86.

4. The reliability of the Moore Test was determined by the test-retest method after a lapse of one week. The correlation coefficients ranged from .95 to .72.

5. Coefficients of correlation between each of the three separate trials were obtained on a group of 441 men and used as partial measures of reliability. Trial one correlated with trial two showed a coefficient of .83; trial two with trial three, .77; and trial one with trial three, .82.

The *Moore Eye-Hand Coordination and Color-Matching Test* is produced and distributed by The California Test Bureau.

AN INVESTIGATION OF A COUNSELOR ATTITUDE QUESTIONNAIRE¹

WILLIAM A. McCLELLAND

Brown University

and

H. WALLACE SINAIKO

New York University

Introduction

ATTITUDES held by a counselor toward his own behavior in counseling situations, and toward various counseling techniques, can have marked implications for effective counseling. However, definitive studies of counselor behavior are practically non-existent. It is the purpose of this study to determine the effectiveness of one technique in the quantitative measurement of these attitudes. Several applications of this technique, the questionnaire method, have been attempted and will be discussed.

Implicit in the questionnaire approach to the investigation of counselor attitudes is the assumption that "correct" responses can be determined. In an investigation of this problem Chase² had 34 judges, "selected because of their known understanding of and ability in counseling," respond to a 101-item Questionnaire he had constructed. Typical items from the Chase Questionnaire are as follows:

- 1 2 3 4 5 Permitting the counselee to express himself freely.
- 1 2 3 4 5 Reprimanding the counselee for displaying aggression.
- 1 2 3 4 5 Advising the counselee to stay on the safe side and not take chances.

The five numbers before each item represent the counselors' attitude toward the practice as follows: 1, Decidedly harmful;

¹This paper was presented at the Midwestern Psychological Association meeting May 8, 1948.

²Chase, Wilton P. "Measurement of Attitudes Toward Counseling." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, VI (1946), 467-473.

2, Probably harmful; 3, Doubtful; 4, Probably good; 5, Decidedly good. A scoring key was developed by counting as "correct" all single ratings of items that received a clear majority of the judges' responses. If two adjacent ratings of an item received a clear majority, both were scored as "correct" responses. Chase was then able to give his Questionnaire to counselor trainees, to compare their responses with those of the 34 judges, and to derive quantitative "scores" (or indices of agreement) with the judges.

The above method of determining "correct" responses in an attitude questionnaire is subject to qualification. First, keying items is undoubtedly some function of the judges' training and experience, their temperaments, and their philosophies of counseling. Thus, it is reasonable to expect considerable variation in keys derived from different groups of judges. Second, the original "set" given the judges might well influence their responses. Chase instructed the judges to "Keep in mind that in every one of the items a general situation is described, and one is therefore not to think in terms of individual cases." But counseling is not in terms of a general situation. It is conceivable that the same counselor dealing with a very dependent student would behave in one way, yet in counseling a how-to-study case might perform quite differently. In short, does the "general counseling situation" exist at all? If the same counselor can have different sets of attitudes for different counseling contacts, then there might be need for as many "correct" questionnaire responses as there are diagnostic categories and counseling philosophies.

For purposes of the present study it was assumed that judges *are* capable of making responses to the Chase Questionnaire items in terms of a general counseling situation. If a meaningful key could be derived, the use of a questionnaire of this type might be of considerable value in the selection and training of counselors. In an agency where counselors specialize in one or more problem areas, inspection of the individual responses and total score would be of value in the placement of counselors. In a counselor-training situation early identification of attitudes at variance with local, empirically defined, "correct" attitudes might facilitate the orientation of the training. These are two possible uses of such an instrument.

Method and Subject

A list of expert counselors, each of whom was to re-evaluate the Chase Questionnaire items, was compiled. The group included only persons who had had at least ten years' counseling experience in and around the University of Minnesota, or who had obtained the Ph.D. degree in Personnel Psychology at that institution. The thirteen expert counselors selected may be characterized as homogeneous by training and counseling experience. Four of the group were academic instructors in counseling courses, four were full-time counselors, and five divided their time between personnel administration and counseling. The judges were given a shortened form of the Chase Questionnaire: ten of the original 74 scorable items were eliminated because they were specific to military separation counseling.

A "Minnesota key" for scoring the questionnaires was obtained as follows: The mean and standard deviation was computed for each of the 64 items. Responses were weighted on the five-point scale described above. Those items were eliminated which had a standard deviation of .8 or larger (arbitrarily selected since these items could have more than two "correct" responses). Inspection of the distributions of judges' ratings supplemented this application of summary statistics. Forty items remained on which there seemed sufficient agreement between the judges so that either one single or two adjacent ratings could be scored as "correct."

The subjects of the investigation were students in counseling courses and counselors at the University of Minnesota. They were administered the 64-item Questionnaire in the spring of 1947 during the first week of the new term. Subjects came from two sources: 106 were students in either of two courses dealing with guidance techniques and counseling practices, and 53 were graduate students in Psychology or Educational Psychology who were either taking graduate courses in Counseling or were engaged in half-time college counseling. The former group of students answered the Questionnaire a second time, at the final session of the course in which they were enrolled. The students' questionnaires were scored with the Minnesota key and with the Chase key for the forty scorable items.

Results

In the first part of the study, summary statistics for the 40 items keyed by the Minnesota judges, and these same items keyed by Chase's judges, showed limited spread. The combined group of students (graduate and undergraduate) had an average score of 30 and a standard deviation of 3 items on the Minnesota key, and an average score of 27 with a standard deviation of 4 on the Chase key. As scores obtained from the two keys correlated $.20 \pm .09$, it would appear that the keys are quite dissimilar. However, inspection reveals two facts: Twenty-four of the 40 items had two adjacent ratings keyed "correct" by Minnesota judges, while only seven of the same 40 items had double ratings on the Chase key. On the Chase key, 32 of the 40 items have one extreme or the other keyed as "correct," while Minnesota judges used the extreme responses for only 24 of the 40 items. These facts, the greater tendency to key adjacent responses as "correct" by Minnesota judges and a reluctance to key extreme values on the part of these judges, could account for the higher mean and smaller variability of the Minnesota scale and possibly for the low correlation between the latter scale and that of Chase.

To test the hypothesis that both keys were really different they were compared in terms of a trichotomy of "good," "harmful," and "doubtful." Under such a comparison only three of the 40 items were classified differently. Three counseling practices rated "good" by Chase's judges were considered of "doubtful" value by the Minnesota raters.

One other possibility was suggested as an explanation for the low inter-scale correlation, namely, the reliability of the 40-item Questionnaire. Reliability estimated in several ways (Kuder-Richardson and split-half uncorrected) turned out to be about .20. To achieve minimum satisfactory reliability the number of items would have to be increased fivefold. Assuming the appropriateness of these tests of reliability, it appears that the two scales correlate about as highly with each other as single administration reliability allows.

The writers feel that further analysis of such low reliability is not called for. Therefore, in spite of the unreliability of its

principal instrument, this study is presented for whatever interest it may be to the reader.

The second part of the study involved an assessment of the effects of instruction upon counseling attitudes. The Questionnaire was administered both before and after the undergraduate courses were given. In these two classes the mean scores obtained with the Minnesota key were about 30, with a standard deviation from 3 to 4 items on the pre-course administrations. In one class there was a slight, but not statistically significant, movement of the post-course mean score upwards toward the score of the instructor. In this class the post-course score correlated $.11 \pm .25$ with final course grades. In the second class the post-course mean score was significantly *lower* (C.R. = 2.8) than that group's pre-course mean score. Just why there should have been movement away from the instructor's score in this second class is not readily apparent. It may be that this instructor was somewhat inconsistent in answering the Questionnaire and in his actual teaching practices; or, simply, that the scale itself is too unreliable. In this second group there appeared to be a moderate degree of relationship between post-course score and course grades ($r = .42 \pm .10$). In any event these data offer equivocal evidence in support of the hypothesis that counseling attitudes can be modified by training, although it is clearer that subject-matter examinations in these two courses are not satisfactory measures of those attitudes. Further use of the Questionnaire with control groups would be helpful.

A third estimate of reliability (which suggests the earlier two are underestimates) is offered by the correlation between pre-course and post-course scores for the 106 undergraduate students, $r = .52$. The amount of time elapsed during the courses was nine weeks.

The final problem investigated was the relationship of the amount of training and experience in counseling to scores on the Questionnaire. The two groups which were compared were the 106 undergraduates and 53 graduate students. The Minnesota key mean-raw-score difference between the two groups is statistically significant (C.R. = 2.8), with the graduates getting the higher scores. This is evidence for the common-sense hy-

pothesis that the longer a student studies in a particular school and/or discipline, the more likely he is to acquire the attitudes of his instructors. Whether or not this greater agreement with the judges is a result of instruction, extra-curricular reading, or personal counseling experience, cannot be answered from these data.

Interpretation and Conclusions

1. Although it was possible to obtain considerable agreement among a carefully selected group of judges on the desirability of certain counseling practices, two obvious limitations of the questionnaire approach to the measurement of counselor attitudes must be mentioned. First, most of the judges spoke about the artificiality of rating practices in a "general counseling situation." They reported that specification of the type of client problem, as well as the nature of the agency function, seemed important in keying "correct" responses. Second, the low reliability of the scale makes the current approach suspect. Perhaps more rigorous item construction and analysis might yield more consistent results.

2. There is equivocal evidence that students' attitudes toward counseling practices are susceptible to change with formal course training, and they are not markedly related to grades in counseling courses.

3. Scores on a scale of counselor attitudes may have some value in differentiating the more-experienced from the less-experienced counselors or trainees in terms of a given set of "correct" responses that have been empirically derived for a local situation.

4. The reservations attendant to the use of the Chase items about counselor attitudes are such as to indicate that they should not be used in their present form. While the approach has possibilities, this study suggests the questionnaire analysis of counselor attitudes requires considerably further investigation before it can be accepted as a useful, reliable tool.

A NOTE ON THURSTONE'S METHOD OF COMPUTING THE INVERSE OF A MATRIX

WILLIAM C. COTTLE

University of Kansas

A RESEARCH worker seeking a concise method of computing the inverse of a matrix will find this in Thurstone's method.

TABLE I

*Computations for Column II, Section B, of Thurstone's Example for Computing the Inverse of a Matrix**

a_{ji}	$-c_{nbj}$	b_{ji}
.48	$-(.60) (.80) = -.48$.000
.80	$-(.60) (.48) = -.288$.512
.36	$-(.60) (.36) = -.216$.144
$C_k 1.64$	$-(.60) (1.64) = -.984$.656
$\Sigma_1 1.64$	$-.984$.656
.00	$-(.60) (1.00) = -.60$	-.600
1.00	$-(.60) (.0) = .00$	1.000
.00	$-(.60) (.0) = .00$.000
$C_k 1.00$	$-(.60) (1.00) = -.60$.400
$\Sigma_2 1.00$	$-.60$.400
$C_k 2.64$	$-(.60) (2.64) = -1.584$	1.056
$\Sigma_3 2.64$	-1.584	1.056

* Not rounded to significant figures as in Thurstone's example (1, p. 47).

The method is applicable to a matrix of any size. Such a person may find also, upon examining this method as outlined by Thurstone, that it would appear to be a rather esoteric solution.¹ Possibly, because of Thurstone's familiarity with the method, he overestimates the cognitive powers of his readers. Clarification of one step in the process of computing the inverse would simplify the method. It is for this purpose that this paper has been written.

¹ Thurstone, L. L. *Multiple Factor Analysis*. Chicago: University of Chicago Press, 1947, pp. 46-48.

TABLE 2
Computations for Column III, Section B, of Thurstone's Example for Computing the Inverse of a Matrix*

a_{ij}	$caib_{ij}$	$caib_{ij}$	$\sum (caib_{ij}) + (caib_{ij})$	b_{ij}
.36	$-(.45) (.80) = -.360$	$-(.27) (0) = .00000$	$-.36000$.00000
.36	$-(.45) (.48) = -.216$	$-(.27) (.512) = -.13824$	$-.35424$.00576
.86	$-(.45) (.36) = -.162$	$-(.27) (.144) = -.03888$	$-.20088$.65912
Ch 1.58 Σ_1 1.58	$-(.45) (1.64) = -.738$ $= -.738$	$-(.27) (.656) = -.17712$ $= -.17712$	$-.91512$ $-.91512$.66488 .66488
.00	$-(.45) (1.00) = -.450$	$-(.27) (-.60) = .16200$	$-.28800$	-.28800
.00	$-(.45) (0) = .000$	$-(.27) (1.00) = -.27000$	$-.27000$	-.27000
1.00	$-(.45) (0) = .000$	$-(.27) (0) = .00000$.00000	1.00000
Ch 1.00 Σ_2 1.00	$-(.45) (1.00) = -.450$ $= -.450$	$-(.27) (40) = -.10800$ $= -.10800$	$-.55800$ $-.55800$.44200 .44200
Ch 2.58 Σ_3 2.58	$-(.45) (2.64) = -1.188$ $= -1.188$	$-(.27) (1.056) = -.28512$ $= -.28512$	-1.47312 -1.47312	1.10688 1.10688

* Not rounded to significant figures as in Thurstone's example.

Thurstone's directions are simple to follow in setting up Section A of the example he gives.² To compute Section B and Section C, the steps are as follows:

1. Column I of Section B is copied exactly from Column I of Section A for both matrices.
2. The reciprocal, $1/b_{11}$, is the reciprocal of the first entry in Column I of Section B, or $1/.80 = 1.25$. This is recorded at the bottom of this column as shown in Thurstone's example.
3. Values for the first column of Section C are computed by multiplying Column I of Section B by this reciprocal,

$$c_{11} = b_{11}(1/b_{11}).$$

(The reciprocal is constant for both the values of the original matrix and those of the identity matrix.)

4. Column II of Section B is computed by the formula:

$$b_{12} = a_{12} - (b_{11}c_{21})$$

where $c_{21} = .60$ is constant for the entire column, both the original matrix and the identity matrix. Computations for Column II of Section B are shown in detail in Table 1.

5. The second column of Section C is computed by the formula:

$$c_{12} = b_{12}(1/b_{22}).$$

Where b_{22} is the second entry in Column II of Section B.

6. Column III of Section B is computed by the formula:

$$b_{13} = a_{13} - b_{11}c_{31} - b_{12}c_{32}$$

where $c_{31} = .45$ and $c_{32} = .27$ are constants for each appropriate column of Section B. Computations for Column III of Section B are shown in detail in Table 2.

7. The rest of the computations can be followed from Thurstone's explanation.

It is hoped that this explanation will enable anyone to follow this method of computing an inverse. The writer has used Tucker's method,³ and this method in computing an inverse of a matrix of the order of 8×8 , and prefers the latter method. He would suggest also that no rounding of figures be done until the computation of the inverse M^{-1} is reached.

² The writer is indebted to Dr. Clyde Coombs of the University of Michigan for the information necessary to follow Thurstone's explanation. The writer spent three days in company of two competent mathematicians in an unsuccessful attempt to follow the method before resorting to a letter to Dr. Coombs.

³ Tucker, L. "A Method for Finding the Inverse of a Matrix." *Psychometrika*, III (1938), 189-197.

NOMOGRAPH OF PETERS AND VAN VOORHIS' APPROXIMATION FORMULA FOR CORRECTING INTERFUNCTION CORRELATION COEFFICIENTS FOR HETEROGENEITY

WILLIAM A. REYNOLDS
National Broadcasting Company

IN setting up a testing procedure for selection and placement of employees in a large organization, it is often the practice to administer one or two tests to certain groups of applicants, and to add new tests to the schedule from time to time. Thus, when it is desired to construct a test battery from the results of a multiple-regression study, it is found that the populations to which the individual tests have been administered are larger than the population to which the two or more tests in combination have been administered. The larger populations usually have larger standard deviations; statistically they are more heterogeneous. Since they are the populations on which later test batteries will be validated, the information on heterogeneity may be used to predict more accurately the true relationships between two tests which have been administered to but a fraction of the number to which each separately has been administered.

It is well known that the size of a coefficient of correlation is affected by the heterogeneity (range of talent) of the population on which it is computed. If a formula were available for correcting the correlation between two tests by taking into consideration the ranges of talent on both tests, a better estimate of the true correlation between them could be obtained. Such a formula is available in Peters and Van Voorhis' *Statistical Procedures and Their Mathematical Bases*.¹

These authors develop their formula by considering first the problem of estimating a corrected reliability coefficient. The

¹ Peters, C. C. and Van Voorhis, W. R. *Statistical Procedures and Their Mathematical Bases*. New York: McGraw-Hill Book Co., Inc., 1940. Pp. 208-210.

formula for the correction of a reliability coefficient for heterogeneity is given by Peters and Van Voorhis in formula 129, as follows:

$$\frac{\sigma_x}{\Sigma_x} = \frac{\sqrt{1 - R_{11}}}{\sqrt{1 - r_{11}}} \quad \text{(Formula for correcting a reliability coefficient for heterogeneity)} \quad (129)$$

where,

σ_x = standard deviation of the shorter range of talent

Σ_x = standard deviation of the longer range of talent

r_{11} = reliability coefficient of the shorter range of talent

R_{11} = reliability coefficient of the longer range of talent.

Any unknown term of this formula may be easily computed by nomograph 55 in the *Handbook of Statistical Nomographs* by Dunlap and Kurtz.²

The case of *inter-function correlation* is more complicated. The assumption is made that the "variance of the distribution of true scores in the one function from their corresponding true scores in the other function is the same in the shorter range as it is in the longer one." The following formulas are derived:

$$\frac{\sigma_x}{\Sigma_x} = \frac{\sqrt{R_{11x} - (R_{12y}^2/R_{11y})}}{\sqrt{r_{11x} - (r_{12y}^2/r_{11y})}}; \quad \text{and} \quad (130)$$

$$\frac{\sigma_y}{\Sigma_y} = \frac{\sqrt{R_{11y} - (R_{12x}^2/R_{11x})}}{\sqrt{r_{11y} - (r_{12x}^2/r_{11x})}}$$

(Formula for correcting inter-function r 's for heterogeneity)

Since these formulas involve reliability coefficients of the measurements in both functions for both ranges, and information regarding these usually is not available, a formula using obtained scores rather than true scores is presented:

$$\frac{\sigma_x}{\Sigma_x} = \frac{\sqrt{1 - R_{12y}^2}}{\sqrt{1 - r_{12y}^2}} \quad \text{(Approximate formula for correcting inter-function } r\text{'s for heterogeneity)} \quad (131a)$$

Similarly,

$$\frac{\sigma_y}{\Sigma_y} = \frac{\sqrt{1 - R_{12x}^2}}{\sqrt{1 - r_{12x}^2}} \quad (131b)$$

² Dunlap, J. W., and Kurtz, A. K. *Handbook of Statistical Nomographs*. Yonkers-on-the-Hudson: World Book Co., 1932.

where the assumption is made that the standard errors of estimate are the same in the shorter range as in the longer one.

Under the condition of having an r_{xy} between two tests in a shorter range of talent than the range of either test taken separately, the r_{xy} must be corrected twice: once for heterogeneity in each test. Taking first the correction for heterogeneity in the x variable, formula 131a may be expressed as follows:

$$R_{xy} = \sqrt{1 - \frac{\sigma_x^2(1 - r_{xy}^2)}{(\Sigma_x)^2}} \quad \begin{array}{l} \text{(Approximate correlation coefficient when the } x \text{ variable has} \\ \text{been corrected for heterogeneity)} \end{array} \quad (A)$$

In turn, correcting R_{xy} in formula (A) for heterogeneity in the y variable, we get:

$$R'_{xy} = \sqrt{1 - \frac{\sigma_y^2(1 - R_{xy}^2)}{(\Sigma_y)^2}} \quad \begin{array}{l} \text{(Approximate correlation coefficient when both the } x \text{ and } y \text{ variables} \\ \text{have been corrected for heterogeneity)} \end{array} \quad (B)$$

Where R'_{xy} is the corrected coefficient of correlation when both variables are corrected for heterogeneity, and R_{xy}^2 is the coefficient obtained from formula (A).

The solution of these equations is rather involved but can be estimated quite simply on the accompanying nomograph. The steps to be taken may be illustrated on the following problem.

From the shorter range of talent, the group to which tests x and y both were administered, the correlation coefficient and the standard deviations of x and y were found to be

$$r_{xy} = .65 \quad \sigma_x = 12.0 \quad \sigma_y = 18.0$$

And from the longer range of talent, the whole populations on which either of the tests were administered, the standard deviations were found to be

$$\Sigma_x = 15.0 \quad \Sigma_y = 22.0$$

The problem is to find:

R_{xy} = coefficient of correlation corrected for heterogeneity in x , and

R'_{xy} = coefficient of correlation corrected for heterogeneity in both x and y .

Step 1. Place a straightedge on the line at the left which corresponds to the standard deviation obtained on the shorter

range of talent in the x variable ($\sigma_s = \sigma_x = 12$) across to the correlation coefficient $r_s = r_{xy} = .65$. (The subscript s on the nomograph refers to "shorter" or sample distribution, although this standard deviation may occasionally be greater than that of the "longer" distribution of the whole population to which the test has been administered.)

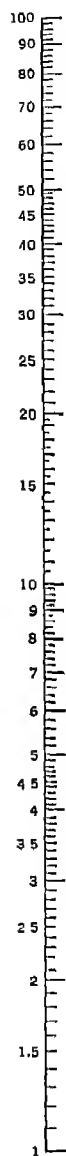
Step 2. Place a pin on the middle reference line of the nomograph, and pivot the straightedge so that it reads the standard deviation obtained from the longer distribution ($\sigma_L = \Sigma_x = 15$) of test x , and read off the corrected coefficient ($r_L = R_{xy} = .80$) from the scale at the right. This obtains R_{xy} , the correction for r_{xy} , when the x variable alone has been corrected for heterogeneity.

Step 3. In turn to correct the R_{xy} for heterogeneity in the y variable, place a pin on the value for the corrected coefficient ($R_{xy} = .80$) and pivot the straightedge so that it reads the standard deviation of test y on the shorter range of talent ($\sigma_s = \sigma_y = 18$).

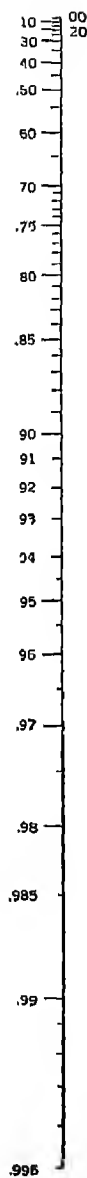
Step 4. Pivot again with a pin held on the middle reference line; change the left side of the straightedge to read the standard deviation on the longer range of talent of test y ($\sigma_L = \Sigma_y = 22$). The result, $R'_{xy} = .87$, is read from the scale at the right. This is the correlation coefficient between the two tests when both have been corrected for heterogeneity.

Quite often on a matrix of intercorrelations such as would be obtained in the development of a test battery of aptitude tests, the inter-function correlation coefficients will be corrected upward when the correction for heterogeneity is made in one variable, but will be reduced when corrected again for the heterogeneity in the other variable. When this occurs, it will be caused by the standard deviation of the test group on one variable being larger than the standard deviation reported in the published norms or obtained on the larger industrial group to which the test has been administered. But in the cases where the standard deviations of the two restricted distributions are less than the standard deviations of the corresponding wider distributions, the corrected correlation coefficients always will be higher.

σ_s OR σ_L



r_s OR r_L



Approximate Formula for Correcting Inter-function r 's for Heterogeneity

A SINGLE CHART FOR TETRACHORIC r

WILLIAM LEROY JENKINS

Lehigh University

THE widely-used Thurstone diagrams¹ for determining tetrachoric r are now out of print. As a substitute, a short-cut method has been devised which employs a single chart.

Essentially, the chart compares the actual percentage-in-excess-of-chance in one cell with the percentages-in-excess-of-chance for r 's of .90, .80, .70, and .60. The interpolation is made graphically if the r is above .60 and arithmetically if the r is lower.

Method with Examples

$$\text{Example A: } \begin{array}{c|c} 273 & 296^* \\ \hline 423 & 24 \end{array}$$

$$\text{Example B: } \begin{array}{c|c} 55 & 80 \\ \hline *70 & 20 \end{array}$$

1. Mark the number (*) in the upper right or lower, left whichever number is smaller.

2. Compute the two percentiles at which the distributions are cut.

$$\frac{296 + 273}{1016} = 56.1\% \text{ (above)}$$

$$\frac{70 + 20}{225} = 40.0\% \text{ (below)}$$

$$\frac{296 + 24}{1016} = 31.5\% \text{ (right)}$$

$$\frac{70 + 55}{225} = 55.6\% \text{ (left)}$$

3. Multiply the two percentiles to obtain the chance percentage in the marked cell.

$$17.6\%$$

$$22.2\%$$

4. Compute the actual percentage in the marked cell.

$$\frac{296}{1016} = 29.2\%$$

$$\frac{70}{225} = 31.1\%$$

5. Subtract result 3 from result 4 to obtain the actual per-

¹ Chesire, L., Saffir, M. and Thurstone, L. L. *Computing Diagrams for the Tetrachoric Correlation Coefficient*. Chicago: Chicago University Press, 1933.

centage in excess of chance. Draw a vertical line downward from this value on the scale at the top of Figure I.

11.6%

8.9%

6. In *each* of the four sets of curves in Figure I: Find the larger cutting percentile on the ordinate scale. Move across in-

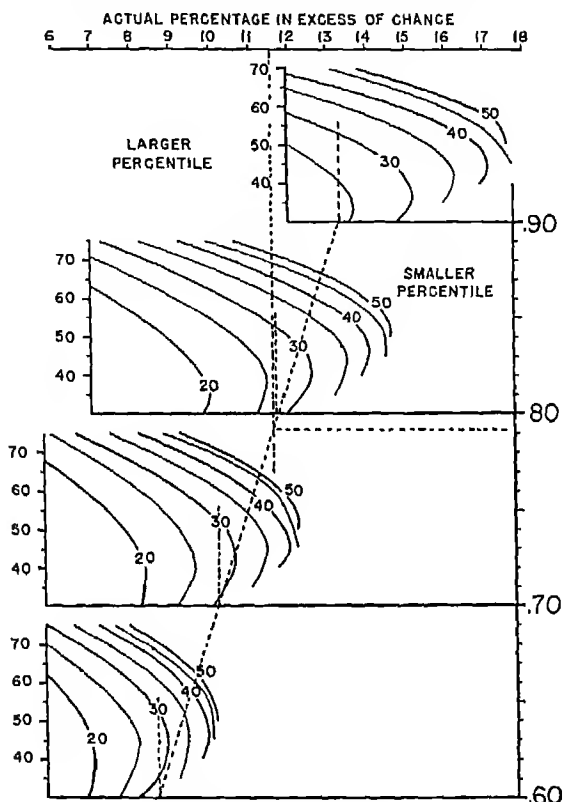


FIG. I. Tetrachoric Chart

The dotted lines indicate the solution for Example A in the text.

terpolating between the curves to the smaller cutting percentile. Drop a vertical to the baseline and mark this point. (The four points represent respectively the percentages-in-excess-of-chance for r 's of .90, .80, .70, and .60.)

7. Through the points marked on the four baselines, draw a curve. Where the curve intersects the straight line drawn in step 5, read off the tetrachoric r from the scale on the right.

tetrachoric $r = .89$
(Example A)

8. If the curve does not intersect the vertical line, the tetrachoric r is less than .60. Make an arithmetical interpolation as indicated below.

$$\text{Tetrachoric } r = .60 \times \frac{8.9}{9.6} = .56$$

(Example B)

The chart in Figure I is too small for actual use, but the author will be glad to furnish without charge a photoprint reproduction of an $8\frac{1}{2} \times 11$ chart on cross-section paper. In empirical tests this chart appears to give the same answers as the Thurstone diagrams.

NEW TESTS*

Algebra Prognosis Test, by Corydon L. Rich. Designed as a help in forecasting a pupil's work and as a guide for sectioning when numbers warrant more than one section. Range: high school and college. Working time: 32 minutes. Published by the C. A. Gregory Co.

Aptitude Tests for Occupations by Wesley S. Roeder and Herbert B. Graham. There are six tests in the battery attempting to measure personal-social, mechanical, general sales, clerical routine, computational and scientific aptitudes. Range: high school and college students, and adults. Working time: 1 hour and 50 minutes for complete battery. Published by California Test Bureau.

Children's Apperception Test, by Leopold Bellak. A personality test specifically designed for use with children between three and ten years of age, of both sexes and of all ethnic groups. Consists of ten pictures of animals in various social situations. Price: set of pictures, manuals and 30 record analysis blanks, \$9.00. Published by C. P. S. Company.

Comprehensive Examination in Psychology, by M. Pullins Claytor. An achievement test for college students in psychology. Working time: 50 minutes. Published by the C. A. Gregory Co.

Cooperative General Culture Test (Forms X and Y), by Norman J. Blair, Jeanne M. Bradford, Mirian May Bryan, Paul J. Burke and Herbert Danzer. Designed to provide an indication of the student's general cultural background. The content has been determined by the consensus of a number of scholars in various fields. Consists of six sections covering current social problems, history and social studies, literature, science, fine arts and mathematics. Range: college students. Working time: 180 minutes. Price: test booklets, per package of 25, \$3.90; answer sheets, per package of 25, \$1.70. Published by Cooperative Test Division of the Educational Testing Services.

*The tests listed bear 1949 or 1950 copyright dates. The addresses of publishers are given at the end of section. In some instances, certain details (particularly prices), are not included because they were not available at the time of going to press.

Cowan Adolescent Adjustment Analyzer, by Edwina A. Cowan, Wilbert J. Mueller and Edna Weathers. Intended for use as a screening device to discover individuals who would profit from referral to visiting teachers, psychiatrists, guidance counselors, etc. Range: junior and senior high school. Working time: no time limit. Price: \$2.65 per package of 25 tests. Published by Bureau of Educational Measurements, Kansas State Teachers College.

Diagnostic Tests of Achievement in Music (Form A), by M. Lela Kotick and T. L. Torgerson. Enables the teacher to determine each pupil's level of mastery of the basic theory and skills in music and to locate the nature of the weaknesses or difficulties in music fundamentals for individuals as well as classes. Range: school music classes. Working time: approximately 45 minutes. Published by California Test Bureau.

Geometry Attainment Test, by R. D. Walton. An achievement test for students with 6 months or more of geometry. Working time: 90 minutes. Price: tests 5/- per dozen; manual, 1/- each. Published by University of London Press, Ltd.

Graded Arithmetic-Mathematics Test, by Philip E. Vernon. Constructed, like the Stanford-Binet intelligence scale, from sets of short problems, one set for each year level. Scores are expressed in Arithmetic-Mathematics Ages from 7-21 years. Range: ages 7-21. Working time: 20 minutes. Published by University of London Press, Ltd.

Guilford-Zimmerman Temperament Survey, by J. P. Guilford and Wayne S. Zimmerman. Scores are obtained for the following areas: general activity, restraint, ascendance, sociability, emotional stability, objectivity, friendliness, thoughtfulness, personal relations and masculinity. Range: senior high school, college and adults. Working time: approximately 45 minutes. Price: package of 25 reusable answer booklets, \$3.75; answer sheets, 3¢ each. Published by Sheridan Supply Company.

Heston Personal Adjustment Inventory, by Joseph C. Heston. Designed to measure, for guidance purposes, the personal adjustment of the normal individual in six areas: analytical thinking, sociability, emotional stability, confidence, personal relations, home satisfaction. Range: high school and college. Working time: 40 to 50 minutes (no time limit). Price: \$2.25

per package of 25 tests. Published by World Book Company.

Holborn Vocabulary Test for Young Children, by A. L. Watts. Consists of 100 questions concerning body parts, household articles, eating, drinking, actions with hands and fingers, etc., to be answered orally by the child. Range: $3\frac{1}{2}$ years of age and upward. Working time: no time limit. Price: 1/-. Published by George G. Harrap and Company, Ltd.

Iowa Every-Pupil Tests of Basic Skills (Form 1), prepared under the direction of E. L. Lindquist. A battery of tests designed to measure certain skills involved in reading, work-study, language and arithmetic at the elementary-school level. There are fourteen separate tests: *Reading Comprehension, Vocabulary, Map Reading, Use of References, Use of Index, Use of Dictionary, Graphs, Punctuation, Capitalization, Language Usage, Spelling, Arithmetic Concepts, Arithmetic Processes and Arithmetic Reasoning*. Range: grades 5-9. Working time: five and one-half hours for complete battery. Price: available upon application. Published by Science Research Associates.

Metropolitan Readiness Tests, by Gertrude H. Hildreth and Nellie L. Griffiths. Consists of six subtests designed to measure a child's readiness to undertake the work of the first grade. The first four tests measure comprehension of words and sentences and visual perception, the fifth measures number knowledge and the sixth measures a combination of visual perception and motor control. Contains also a supplementary Drawing A Man test. Range: pre-first grade children. Working time: approximately 60 minutes. Price: \$2.10 per package of 25 tests. Published by World Book Company.

Murphy-Durrell Diagnostic Reading Readiness Test, by Helen A. Murphy and Donald D. Durrell. Designed to furnish measure of three critical abilities: auditory discrimination, visual discrimination, learning rate. It is a test for group use. Range: intended for first graders. Working time: test 1 and 2 approximately 1 hour (no time limit); learning rate test has specific time limits. Price: \$1.55 per package of 25 test booklets; \$1.25 per package of flash cards. Published by World Book Company.

Musical Aptitude Test (Series A), by Harvey S. Whistler and Louis P. Thorpe. Designed to measure an individual's aptitude for the study of music. Consists of five parts: rhythm recogni-

tion, pitch recognition, melody recognition, pitch discrimination, and advanced rhythm recognition. Range: grades 4-10. Working time: approximately 40 minutes. Published by California Test Bureau.

Revere Safety Test, by Revere Copper and Brass, Inc., in cooperation with the Psychological Evaluation and Services Center, Syracuse University. Designed to measure knowledge of correct safety procedures in the industrial situation. The subject is required to tell whether each of 162 pictures illustrates good or bad safety practices. Four areas of industrial safety are covered: general safety, pilings, carrying and traffic, tools and machine operation. Working time: 20 minutes. Price: reusable test booklets, each 30¢; answer sheets, package of 25, \$1.00; scoring stencil, each 10¢. Published by Science Research Associates.

Small Parts Dexterity Test, by John F. and Dorothea M. Crawford. A performance test designed to measure fine eye-hand coordination. Part I measures dexterity in using tweezers to insert small pins in close-fitting holes in a plate and to place small collars over protruding pins. Part II measures dexterity in placing small screws in threaded holes in a plate and screwing them down with a screwdriver until they drop through the plate into a metal dish below. Working time: about 15 minutes. Price: \$25.00 complete with manual and spare parts. Published by Psychological Corporation.

SRA Self-Scorer, by Maurice E. Troyer and George W. Angell. A new type of answer sheet designed for use with any teacher-constructed objective test. Its primary function is to promote student learning by immediately revealing whether test question has been answered correctly or incorrectly. Questions must be arranged to fit one of the eight answer keys. Four types of answer keys are provided: 1) true-false (space for 300 questions); 2) true-false and multiple choice (space for 210 questions); 3) four-choice (space for 150 questions); and 4) five-choice (space for 150 questions). Each of these types is published in two different forms. Price: self-scorer, complete, each \$1.50; answer sheets, per package of 25, \$1.00. Published by Science Research Associates.

SRA Youth Inventory (Form A), by H. H. Remmers and Benjamin Shimberg. A check list of 298 questions that has been designed as a tool to help teachers, counselors and school administrators to identify quickly the problems that young

people say worry them most. Range: teen-age students
Working time: approximately thirty minutes. Price: reusable
booklet with answer pad, 48¢; package of 25 answer pads,
\$1.75; scoring stencil, 50¢. Machine scored form: reusable
answer pad, 42¢; package of 100 answer sheets, \$2.90;
scoring stencils \$2.50. Published by Science Research As-
sociates.

Social Intelligence Test, by J. A. Moss, T. Hunt and K. A. Omwake.
Designed to measure one's ability to get along with others.
Consists of five parts measuring: judgment in social situa-
tions, memory for names and faces, observation of human
behavior, interpretation of mental state from spoken or
written words, and sense of humor. Range: for high school,
college and industrial use. Working time: 45 minutes (two
shorter forms of 40 minutes and 30 minutes are available).
Price: \$3.75 per package of 25 tests (regular form). Pub-
lished by Center for Psychological Service, George Washing-
ton University.

State High School Testing Service for Indiana offers a list of 49 sub-
ject-matter tests, intelligence scales and inventories based
on the Indiana courses of study, approved text books and
teaching practices. The list is as follows: *Agriculture*: Animal
Husbandry, Farm Shop Tools (Forms A and B); *Commercial*:
Commercial Arithmetic, Bookkeeping (first and third semes-
ters), Shorthand (first and third semesters), Typewriting (first
and third semesters); *English*: Mechanics of Written Eng-
lish (grades 9-12), Tools of Written English (grades 7-8),
Purdue Reading Test (grades 7-12); Health and Safety
Education; *Home Economics (high school)*: Child Develop-
ment, Clothing I (Forms A and B), Clothing II, Foods I,
Foods II, Home Care of the Sick, Housing of the Family;
Home Economics (grades 7-8): Care and Play of Children,
Clothing Problems, Food in the Home, Housekeeping; *Lang-
uage*: French Recognition Vocabulary (Forms K and L),
Latin (first and third semester), Spanish (first semester);
Mathematics: Algebra (first and third semesters), Arithmetic
Fundamentals (Forms A and B), Plane Geometry (first
semester), Solid Geometry, Trigonometry; Mechanical
Drawing; *Science*: Biology (first semester), Chemistry (first
semester), General Sciences (first semester), Physics (first
semester); *Social Studies*: Civics-Junior High School (first
semester), Civics-Senior High School (first semester), Civics-
Senior High School (one semester course), Economics, Amer-
ican History (first semester), World History (first semester),
Two Thousand Test Items in American History (bound,
90¢); *Guidance*: A.C.E. Psychological Examination, Out

Quick-Scoring (Gamma Am), Henmon-Nelson (grades 7-12), High School Attitude Scale (Forms A and B), Purdue Personality Schedule, Maturity Rating Scale, Purdue Physical Science Test; *Teacher Self-Evaluation: A Diagnostic Teacher Rating Scale* (grades 4-8, Forms A and B), Purdue Rating Scale for Instruction (in lots of 500 or more). The prices of these tests range from 1½¢ to 6¢, plus 1¢ per copy for tests going out of state. Exact prices may be obtained from publishers, State High School Testing Service for Indiana, Purdue University, Lafayette, Indiana.

Tests for Infants 4-12 Weeks Old (Test A), by A. R. Gilliland. Designed to measure adaptation to the physical and social environment. Price: \$2.00 per package of 25 test record sheets and examiner's manual. Test equipment may be obtained from the author at Northwestern University. Published by Houghton Mifflin Company.

Test of English Usage (Forms A & B), by Henry D. Rinsland, Raymond W. Pence, Betty S. Beck and Roland L. Beck. Designed to measure the student's ability to recognize and apply the basic rules of English composition. Consists of three parts: mechanics of writing; accurate use of words; building sentences and paragraphs. Range: high school and college. Working time: no time limit. Published by California Test Bureau.

Wechsler Intelligence Scale for Children, by David Wechsler. A psychodiagnostic instrument which has grown logically out of the Wechsler-Bellevue intelligence scales used with adolescents and adults. In fact, most of the items in the W.I.S.C. are from Form II of the earlier scales, the main addition being new items at the easier end of each test to permit examination of children as young as five years of age. Range: primarily for use with the school-age child. Working time: 45 minutes to 1 hour. Price: \$19.50, including manual and 25 record forms. Published by the Psychological Corporation.

ADDRESSES OF THE PUBLISHERS AND DISTRIBUTORS OF THE TESTS LISTED

Bureau of Educational Measurements, Kansas State Teachers College, Emporia, Kansas.
Bureau of Publications, Teachers College, Columbia University, New York City.

- C. P. S. Company, P.O. Box 42, Gracie Station, New York 28, New York.
- California Test Bureau, 5916 Hollywood Boulevard, Hollywood 28, California.
- Center for Psychological Services, George Washington University, Washington 6, D. C.
- Educational Testing Service, 20 Nassau Street, Princeton, N. J.
- The C. A. Gregory Company, 345 Calhoun Street, Cincinnati 19, Ohio.
- George G. Harrap and Company, Ltd., 182 High Holborn, London, WC 1, England.
- Psychological Corporation, 522 Fifth Avenue, New York 18, New York.
- Science Research Associates, 228 S. Wabash Avenue, Chicago 4, Ill.
- Sheridan Supply Company, P.O. Box 837, Beverly Hills, California.
- The Steck Company, 9th at Lavaca, Austin, Texas.
- University of London Press, Ltd., Warwick Square, London, E.C. 4, England.
- World Book Company, Yonkers-on-Hudson, New York.
-

THE CONTRIBUTORS

Anne Anastasi—Ph.D., Columbia University, 1931. Lecturer in Psychology, 1929-1930; Instructor in Psychology, 1930-1939, Barnard College. Assistant Professor and Chairman of the Department, Dept. of Psychology, Queens College, N. Y. City, 1939-1946. Associate Professor of Psychology, Graduate School, Fordham University, 1947-. Author of *Differential Psychology*; co-author of *Fields of Psychology* and *Foundations of Psychology*; author of approximately 50 monographs and articles in psychological journals. Fellow, American Association for the Advancement of Science, American Psychological Association (Divisional Representative, 1947-), New York Academy of Sciences (Chairman, Section of Psychology, 1940). Member, Phi Beta Kappa, Sigma Xi, New York State Psychological Association (Vice-President, 1945), Eastern Psychological Association (Board of Directors, 1944-1946, 1948-; President, 1946-1947).

Donald K. Beckley—Ph.D., University of Chicago, 1948. Instructor, Rochester (New York) Institute of Technology, 1939-1942. Staff member, Examination Staff for the United States Armed Forces Institute, University of Chicago, 1942-1943. Professor of Retailing and Director, Simmons College Prince School of Retailing, 1946-. Co-author of *Merchandising Techniques* and *The Retail Sales-person a Work*, and author of articles on employment testing in retailing.

Walter R. Borg—Ph.D., University of California, 1948. Assistant in Educational Psychology, University of California, 1946-1948. Assistant Professor of Educational Psychology, University of Texas, 1948-.

Clyde H. Coombs—Ph.D., University of Chicago, 1940. Research Assistant, Psychometric Laboratory, 1937-1940. Research Assistant, Mathematical Biophysics, University of Chicago, 1940-1941. Personnel Psychologist, War Dept., 1941-1946. Assistant Professor of Psychology, 1947-1948, Associate Professor, 1948-, University of Michigan. Fellow, American Psychological Association. Member, Psychometric Society, American Statistical Association, Institute of Mathematical Statistics, Phi Beta Kappa, Sigma Xi.

Lee J. Cronbach—Ph.D., University of Chicago, 1940. Instructor, Assistant Professor, Associate Professor, State College of Washington, 1940-1946. Associate Psychologist, University of California Division of War Research, 1944-1945. Assistant Professor of Education, University of Chicago, 1946-1948. Associate Professor of Education, Bureau of Research and Service, University of Illinois, 1948-. Author

of *Essentials of Psychological Testing*, and articles. Fellow, American Psychological Association. Member, American Educational Research Association.

William C. Cottle—Ed.D., Syracuse University, 1949. New York State Public Schools, 1932-1945. Instructor and Chief of Veterans Testing Service, Syracuse University, 1945-1947. Assistant Professor of Education and Counselor, 1947-1948; Associate Professor and Assistant Director, Guidance Bureau, 1948, University of Kansas. Associate Member, American Psychological Association. Professional Member, National Vocational Guidance Association. Member, Phi Delta Kappa, Kansas Psychological Association, Kansas Academy of Science, American Association of University Professors.

Edward E. Cureton—Ph.D., Columbia University. Associate Professor and Professor of Education, Alabama Polytechnic Institute, 1931-1941. Senior Educational Statistician, U. S. Office of Education, 1941-1942. Chief, Testing Unit, Civilian Personnel Branch, Hq. Army Service Forces, 1942-1943. Chief, Civilian Personnel Research Subsection, Adjutant General's Office, 1943-1945. Chief, Technical Operations and Control, Personnel Research Section, Adjutant General's Office, 1946-1947. Staff member, Richardson, Bellows, Henry and Co., Inc., 1945-1946; 1947-1948. Professor of Psychology, University of Tennessee, 1949-. Fellow, American Psychological Association, American Association for the Advancement of Science. Member, American Educational Research Association, Institute of Mathematical Statisticians. Past President, Psychometric Society.

Frank J. Dudek—Ph.D., University of Southern California, 1947. Aviation Psychologist, AAF Aviation Psychology Program, 1942-1946. National Research Council Predoctoral Fellow, 1946-1947. Assistant Professor of Psychology, Northwestern University, 1947-. Associate Member, American Psychological Association. Member, Midwestern Psychological Association, Phi Beta Kappa, Sigma Xi, Phi Delta Kappa, Phi Kappa Phi, Psi Chi, American Association of University Professors.

William Leroy Jenkins—Ph.D., University of Michigan, 1936. Instructor, Assistant Professor, Lehigh University, 1935-43. Research Associate, University of California Division of War Research, 1943-44. Supervisor, Training Aids, Columbia University Division of War Research, Submarine Training Section, 1944-45. Associate Professor of Psychology, Lehigh University, 1946-. Author of articles on cutaneous sensitivity. Member, American Psychological Association.

Robert B. Kamm—Ph.D., University of Minnesota, 1948. Member, Counseling Staff, The General College, University of Minnesota, 1946-1948. Dean of Students, Drake University, at present. Associate Member, American Psychological Association. Member, American College Personnel Association, Phi Delta Kappa.

Robert W. Kleemeier—Ph.D., University of Michigan, 1942. Teaching Fellow, University of Michigan, 1938–1941. Clinical Counselor and Instructor in Psychology, University of Illinois, 1941–1942. Instructor in Psychology, Northwestern University, 1942–1943. Classification and Selection Officer, U. S. Maritime Service, 1943–1945. Assistant Professor of Psychology, Northwestern University, 1946–. Fellow, American Psychological Association. Member, Midwestern Psychological Association, Phi Beta Kappa, Sigma Xi, Phi Sigma, American Association of University Professors.

William A. McClelland—Ph.D., University of Minnesota, 1948. With the U. S. Army Air Forces: Aviation Psychologist, 1942–1946. Instructor, University of Minnesota, 1946–1948. Assistant Professor of Psychology, Brown University, 1948–. Member, Phi Beta Kappa, Sigma Xi, Psi Chi, Phi Delta Kappa, American Psychological Association, National Vocational Guidance Association, American Association for the Advancement of Science, Psychometric Society.

Joseph E. Moore—Ph.D., George Peabody College, 1935. Instructor in Psychology, North Carolina State College, 1931–1935. Professor of Psychology, George Peabody College, 1936–1942. With the U.S. Army: Classification Officer, 1942–1945; Personnel Consultant, 1944–1945. Professor and Chairman of the Department of Psychology, Georgia Institute of Technology, 1945–. Member, American Psychological Association, Diplomate in Industrial Psychology, American Board of Examiners in Professional Psychology, Southern Society of Philosophy and Psychology. President, Georgia Psychological Association, 1948. Professional Member, American Vocational Guidance Association, Phi Kappa Phi, Phi Delta Kappa.

William A. Reynolds—M.A., University of California, 1941. Instructor in Psychology, Bakersfield Junior College, 1941–1942. Psychologist, War Dept., Placement and Testing Branch, McClellan Field, Sacramento, Calif., 1942–1944. Statistical Analyst, War Dept., Wright Field, Dayton, Ohio, 1944–1945. Research Associate, Research Dept., National Broadcasting Co., 1946–1949. Lecturer, New York University, 1947, 1949. Author of articles on testing and personnel placement, statistical methods, radio research. Member, American Psychological Association, New York State Psychological Association, American Association for Public Opinion Research, American Statistical Association, American Marketing Association, Institute of Mathematical Statistics.

Edward A. Rundquist—Ph.D., University of Minnesota, 1932. Assistant Psychologist, Child Guidance Clinic, Minneapolis Public Schools, 1928–1929. Instructor in Research, Institute of Child Welfare, University of Minnesota, 1929–1930. Chief Psychologist, Child Guidance Clinic, Minnesota Public Schools, 1930–1933. Research Fellow, University of Minnesota, 1933–1934. Chief Psychologist,

Child Study Department, Minneapolis Public Schools, 1934-1935. Assistant Director, Psychological Laboratory, Cincinnati Public Schools, 1935-1942. Various positions in Personnel Research Section, Adjutant General's Office, 1942-1946. Assistant Director Personnel Research, Owens-Illinois Glass Company, 1946-1949. Chief, Personnel Evaluation and Criterion Research, Personnel Research Section, Adjutant General's Office, 1949-. Co-author with Raymond F. Sletto of *Personality in the Depression*. Author of articles in various journals. Member, American Psychological Association, Midwestern Psychological Association, Sigma Xi.

H. Wallace Sinaiko—M.A., University of Minnesota, 1947. Graduate Teaching Assistant, Dept. of Psychology, University of Minnesota, 1946-1947. Assistant Employment Manager, L. Bamberger & Co., Newark, N. J., 1947-1949. Graduate student, New York University, 1949-. Research Psychologist with Human Engineering Laboratory, Research Division, College of Engineering, New York University, 1949-. Member, American Psychological Association, Eastern Psychological Association, New Jersey Psychological Association, Psi chi.

Edward A. Suchman—Ph.D., Columbia University, 1947. Research Assistant, Princeton University, 1937-1939. Research Fellow, Rockefeller Foundation, 1939-1940. Research Associate, Columbia University, 1940-1942. Research Associate, Research Branch, War Dept., 1942-1946. Research Associate, Social Science Research Council, 1946-1947. Assistant Professor and Executive Officer, Dept. of Sociology; Associate Director, Social Science Research Center, Cornell University, 1947-. Author of articles in sociological and psychological journals. Co-author of *The American Soldier; Studies in Social Psychology in World War II*. Member, American Psychological Association, American Sociological Society, Sociological Research Association, Association for American Public Opinion Research.

C. Gilbert Wrenn—Ph.D., Stanford University, 1932. Vocational Counselor, Stanford University, 1928-1936. Associate Director, General College; Associate Professor of Educational Psychology, 1936-1938; Professor of Educational Psychology, 1938-, University of Minnesota. Consultant, Student Personnel, Teacher Education Commission of the American Council on Education, 1939-1942. On military leave with the U. S. Armed Forces, Personnel Officer in Bureau of Naval Personnel and Pacific Area, 1942-1946. Associate, American Youth Commission, 1939-1941. Author and co-author of *Student Personnel Problems, Studying Effectively, Aids to Group Guidance, Time on Their Hands*, and numerous journal articles. President, National Vocational Guidance Association. Vice-President, Council of Guidance and Personnel Association, 1946-.



VOLUME TEN, NUMBER TWO, SUMMER, 1950

<i>The Theory and Classification of Criterion Bias.</i> HUBERT E. BROGDEN AND ERWIN K. TAYLOR	159
<i>An Investigation of Two Hypotheses Regarding the Nature of the Spatial-Relations and Visualization Factors.</i> WILLIAM B. MICHAEL, WAYNE S. ZIMMERMAN AND J. P. GUILFORD	187
<i>On the Use of Interactions as "Error Terms" in the Analysis of Variance.</i> ALLEN L. EDWARDS	214
<i>The Objective Measurement of Dynamic Traits.</i> R. B. CATTELL, A. B. HEIST, P. A. HEIST AND R. G. STEWART	224
<i>The Construction and Validation of a Work-Type Auditory Comprehension Reading Test.</i> GEORGE SPACHE	249
<i>Validation and Standardization of the AGO General Mechanical Aptitudes Test for the Selection of Civilian Employees in War Department Installations.</i> ADAM PORUBEN, JR.	254
<i>Three Aids in the Evaluation of the Significance of the Difference Between Percentages.</i> C. H. LAWSHE AND P. C. BAKER	263
<i>A Study of Faking on the Kuder Preference Record.</i> ORRIN H. CROSS	271
<i>Psychological Testing for Immigrants in a Vocational Counseling Agency.</i> BENJAMIN BALINSKY	278
<i>An Investigation of the Personality Traits of Art Students.</i> MARTIN SPIAGGIA	285
<i>The Knowledge of General Education of a Sample of Syracuse University Students as Revealed by the Cooperative General Culture Test and the Time Magazine Current Affairs Test.</i> N. M. DOWNIE, M. E. TROYER AND C. R. PACE	294
<i>The Full-Range Picture Vocabulary Test: II, Selection of Items for Final Scales.</i> ROBERT B. AMMONS AND LEO D. RACHIELE	307
<i>Does Face Validity Exist?</i> SIDNEY ADAMS	320
<i>Administration of the Purdue Pegboard Test to Blind Individuals.</i> JAMES W. CURTIS	329
<i>Evaluating Psychometric Proficiency.</i> FRANK M. DUMAS	332
<i>Interest and Personality Measures of Veteran and Non-Veteran University Freshman Men.</i> KATHERINE K. FASSETT	338
<i>Award in Student Personnel Research.</i> C. GILBERT WRENN	342
<i>Quick Estimation of Multiple R.</i> WILLIAM LEROY JENKINS	346

THE THEORY AND CLASSIFICATION OF CRITERION BIAS

HUBERT E. BROGDEN

and

ERWIN K. TAYLOR

Personnel Research Section, AGO¹

Introduction

IN that area of psychology concerned with the development of tests and other predictive instruments, psychologists have continually emphasized the need for validation. This insistence is sufficiently pronounced to serve as a trade mark of professional psychologists. It is consistent with this insistence upon validation, that the importance of the criterion problem has been widely recognized. This is particularly true of the many psychologists connected with the various testing programs conducted during World War II. However, little attention and less effort have been devoted to a systematic consideration of the problems involved in criterion construction. Publications by Bellows (1), Stuit (13), Toops (15), Viteles (18) and Guilford (9) are among the few dealing particularly with these problems. Any systematic consideration of the problems involved in criterion construction inevitably leads to the problem of bias; to a consideration of the ways in which components which should properly be a part of the criterion are omitted; to the ways in which extraneous components are introduced; and to how distortion of weighting or of scale units occurs.

[T]his paper will attempt a systematic consideration of these problems. A classification of bias will be introduced and related to the steps involved in criterion construction. The more specific problems of bias encountered will then be discussed in relation to this classification system and in relation to the various types of criteria (i.e., production records, ratings, achievement tests, etc.).

¹ The opinions expressed are those of the authors and do not necessarily express the official views of the Department of the Army

Before proceeding further we should like to discuss two points important to the authors' general orientation in attacking criterion problems. The essence of the first point in question may be stated as follows: In seeking to define criterion problems—particularly those of criterion bias—it must be recognized that the objective of criterion construction is subsidiary to that of selecting the most efficient battery of predictors. Prediction instruments are validated for the purpose of picking the best selection battery, assigning appropriate weight to each of its several components, and determining the effectiveness of the battery. The criterion achieves its sole function if it makes these objectives of validation possible. In the development of an industrial selection program, for example, the criterion should give an accurate and unbiased measure of the extent to which individuals in the validation population contribute to or detract from the efficiency of the organization. This may be taken as axiomatic. If so, the emphasis in criterion construction must be in terms of the objectives of the prediction problem.

Criteria differ from predictors in that the former *must* be tested in terms of a concept that we carefully avoid in the latter. In constructing or choosing from among existing predictors, an empirical approach can be, and often is, profitably used. Recourse to previous research results, information based on job analysis, hunches, hypotheses, and intelligent guesses all provided legitimate bases upon which to predicate a potential selection battery. Wrong guesses can be costly in terms of wasted research resources, but they are not misleading since they are put to the empirical test of how well each accomplishes the objectives of the prediction task, i.e., how well each correlates with the criterion.

The criterion, by contrast, can be subjected to no wholly satisfactory empirical test of its adequacy. The criterion must, consequently, be logically justifiable as valid in its own right. The remainder of this paper is predicated on the acceptance of this point of view. Invalid and biased criteria, again in contrast to predictors, cannot be eliminated through empirical demonstration of their inadequacy. Thus, the faulty criterion not only wastes research efforts, but seriously reduces the effectiveness of the final outcome of the program.

For the purpose of this discussion, a biasing factor may be defined as any variable, except errors of measurement and sampling error, producing a deviation of obtained criterion scores from a hypothetical "true" criterion score. It is apparent that this definition is quite general and leads to the consideration of all factors which bear upon the desirability or undesirability of criterion elements and their combination. Of course, the practical consideration which faces the research worker in a "real" situation precludes the complete elimination of all undesirable aspects of criterion construction. Perfection may be approached—it is not likely to be achieved. Nonetheless, to improve his criteria to the point optimal for the conditions under which he is working, the research psychologist must know the importance of different types of bias, the manner in which each will probably affect his results, the proper emphasis to place upon the elimination of those factors producing a distortion of results of indeterminate magnitude, and, finally, the probable effect of bias that cannot be entirely eliminated. It will be shown that different types of biasing factors vary widely in their distortive effect, generally as a function of the degree of their correlation with the members of the predictive battery. Some biasing factors influence the validity coefficients but have little or no effect on estimates of criterion reliability. Others affect both. Still others may alter the apparent reliability of the criterion without seriously influencing the validity.

Classification of Biasing Factors

Imperfections or bias in the criteria may be classified as.

- (1) *Criterion Deficiency*—omission of pertinent elements from the criterion.
- (2) *Criterion Contamination*—introducing extraneous elements into the criterion
- (3) *Criterion Scale Unit Bias*—inequality of scale units in the criterion.
- (4) *Criterion Distortion*—improper weighting in combining criterion elements

The above classification of criterion bias is functional in terms of the steps the authors consider essential to adequate criterion construction. These steps may be indicated as follows:

- (1) Careful analysis of the total situation in which the criterion behavior occurs for the purpose of isolating all sub-criterion variables and obtaining preliminary estimates of their relative importance-- the determination of what is to be measured.
- (2) The construction of procedures and/or scales for the measurement of these elements- determination of how each element is to be measured.
- (3) Development of a procedure for combining these elements into the desired single composite--determination of the relative importance of each element to over-all efficiency.

Criterion deficiency is most apt to occur in the process of determining the variables to be included in the criterion. *Contamination* and *criterion scale-unit bias* are most likely to appear in the process of constructing scales for the measurement of the sub-criterion elements while *criterion distortion* results primarily from faulty methods of combining the criterion elements.

Each of the three steps of criterion construction is necessarily involved, however sketchily, in the development of any criterion. The rationale of our classification of bias is so intimately related to the belief in the need for an explicit plan of construction involving these three steps as to justify further clarification of the implications of each in its relation to bias.

The desirability of establishing the variables important to "success" by observation and job analysis (step 1) before proceeding to scale construction (step 2) and the combination of sub-criterion variables (step 3) deserves special emphasis. From reports of validation studies found in the literature, it may be judged that the usual first step in criterion development is the search for *available* criterion measures. The psychologist employing this procedure very often arrives at a decision as to criterion content that is undesirably influenced by factors of availability. The discovery of several already available or readily obtained measures that are *apparently* suitable is inclined to lead to neglect of the systematic observation and analysis necessary to insure that all important aspects of on-the-job productivity have been identified. In choosing criteria on the basis of availability, method of measurement as well as

nature of variables usually is also a function of convenience rather than of desirability. Without accomplishing step 1 before deciding upon the means by which the criterion variables are to be measured, a systematic consideration of alternate methods of scale construction or measurement and choice of the optimal method for each criterion variable is not likely to be made. While it is recognized that, in many cases, the final decision as to the method of measurement will have to be made in the light of economy and available research resources, it is the firm belief of the authors that there is generally enough freedom of choice within the limitations imposed by even a policy of strict expediency, to justify the type of analysis proposed. At least the decision can be made with full and explicit recognition of the basis for making it. It might be added, parenthetically, the careful accomplishment of step 1, in addition to insuring that adequacy of criterion variables, frequently serves the additional function of supplying valuable clues as to possible predictors. Savings realized through this means may in part, if not entirely, offset the extra cost and effort required to make a thorough observation and analysis.

Criterion Bias and Predictor Correlation

To this point, our classification and discussion of bias have been in terms of the criterion alone. Since effort expended in constructing a bias-free criterion is, as we have stressed before, directed ultimately toward the proper choice and weighting of a battery of predictors, it is essential to consider the effect of criterion bias on the degree to which this objective is realized.

Biassing factors correlating with the predictors will obviously distort the validities and the partial regression weights of the various predictors. They may even result in the inclusion of tests in the battery that predict only bias and have no relationship to the "true" criterion. The introduction of bias having no relation to the predictors is, on the other hand, equivalent, in effect, to an increase in the error of measurement of the criterion. The relationship of all predictors to the criterion will be attenuated. But this attenuation will be proportional for all predictors. Consequently, the relative magnitude of the validities and the partial regression coefficients will be unaffected.

This leads to the highly important conclusion that the "true" validity of the weighted composite resulting from the validation study remains substantially unaffected by test-free bias, even though the exact magnitude of this validity cannot be estimated. With these considerations in mind, we may further classify biasing factors into those which are *predictor correlated* and those which are *predictor free*.

The authors do not wish to imply that the attenuating effect of test-free bias is of little import. In addition to the attenuation of the validity coefficients and partial regression weights, two other undesirable results will accrue from the introduction of test-free bias into the criterion: (1) The sampling error of the validity and regression weights will tend to increase, thus rendering these statistics less stable from sample to sample, and (2) biasing factors that are test free may, none the less, distort estimates of the reliability of the criterion in an indeterminate manner.

The first of these faults may be overcome by increasing the size of the experimental population if additional cases are available with, of course, a resulting increase in the cost of the research. The problem of correcting for the unknown effect of test-free bias on criterion reliability is more difficult, and possible solutions are usually less satisfactory. Such possible solutions are, in any event, particular to the nature of the biasing factors.

In spite of these adverse effects of test-free bias, it is believed that, effectively, it is the presence or absence of test-correlated bias that "makes" or "breaks" the criterion.

Criterion Deficiency

Before beginning our discussion of criterion deficiency, a distinction should be made between criteria designed to measure over-all proficiency on a particular job and those concerned with success in specific job elements. The validation problems involved are both legitimate. In the latter case, it may be desired to measure success in a job element common to a wide variety of job classifications in order to validate a test designed specifically to predict this element. Adequate validation samples can sometimes be obtained only by combining groups from

a wide variety of jobs, all of which share the concerned element. The problem of criterion deficiency would not usually be pertinent to validation studies of this nature. Our concern, in any event, will be exclusively with criterion deficiency as it occurs in criteria of general on-the-job success.

Criterion deficiency is present to a greater or less degree in all studies involving criteria of general success. While it is doubtful that a criterion could be built which would take into account *all* aspects of on-the-job performance, it is the authors' opinion that the high incidence of deficiency may be avoided by a more systematic approach to the problem of determining criterion elements. In the light of our earlier discussion of the relationship between biasing factors and the steps essential to criterion construction, it is apparent that it is in step 1—the analysis of the situation in which the criterion behavior occurs—that criterion deficiency is most likely to materialize. Adaptation of the principles of worker analysis can probably be made so as to minimize criterion deficiency in prediction problems.

The systematic investigation of the situation in which the criterion behavior occurs serves several valuable functions. First, it minimizes the possibility of overlooking important criterion elements. Second, it supplies the investigator with valuable clues as to the most practical means of measuring the several criterion elements. Third, the analysis supplies some initial estimates of the relative importance of the several criterion elements. Thus, if available facilities require limitation of the criterion to a bare minimum, an intelligent judgment may be made as to which elements may be omitted from the study with least harm. Finally, an analysis of the criterion situation in advance of any other steps in the study will generally shed considerable light on the nature of the predictors most likely to be valid. This can eliminate considerable loss of valuable testing time and may result in batteries of greater validity than would usually be the case with predictors chosen on a less sound basis.

The "critical incident" technique for the construction of rating scales as expounded by Flanagan (7) appears to offer promise as a means of reducing criterion deficiency in rating. Not enough is yet known concerning the use of the method to permit a considered judgment of its value for this purpose.

One factor frequently making for criterion deficiency is the inclination of investigators to employ only one type of criterion measure. Studies using ratings, usually use only ratings; those in which production records are used, use only production records; where job samples are employed, neither ratings nor production records are likely to enter into the picture. If an adequate analysis of the job situation were accomplished and a decision as to criterion content were made before consideration is given to the most desirable measuring techniques for each job element, it would seem that production records would often be found most desirable for some of the criterion elements and ratings or job samples most desirable for other elements.

Composite criteria consisting of a variety of production indexes seem, in practice, to be most frequently and most obviously subject to criterion deficiency. The difficulties involved in devising and putting into operation the procedures necessary to obtain production records for those job elements for which none already exist often constitute the determining factor in such instances of criterion deficiency. A systematic approach to criterion construction will do much to minimize such bias. If the important job elements influencing over-all efficiency are isolated first of all, gaps in the total job picture become more readily apparent and measures may be obtained of those elements necessary to complete the criterion composite in the manner that is most practical in the particular situation. If it is found at that time that production records cannot be made available for the measurement of all criterion elements; ratings, job samples, or other means may be devised to eliminate the gaps in the composite.

In considering criterion deficiency in relation to rating criteria, we must distinguish between over-all ratings and composites derived from separate evaluations for each element. In the latter case, there is the same need for systematic analysis of the job situation for the determination of the elements to be evaluated as in the construction of production record or mixed criteria. Generally, rating criteria, whether as separate element ratings or as over-all, undertake to account for a larger part of the total job than is the case with production criteria. Thus, criterion deficiency is probably somewhat less prominent in

ratings than in ordinary production record criteria. Bias undoubtedly does occur because of improper weighting. It should be pointed out that so little weight is given to some factors that the criterion distortion introduced practically amounts to criterion deficiency.

It should be recognized that in the use of over-all ratings of effectiveness, the problem of criterion deficiency has not been solved. Rather, it has been placed in the laps of the raters. The extent to which such rating will be deficient depends, of course, upon the extent to which each of the raters has included each of the important elements of success in making his rating. It may be expected that different raters will incorporate different elements into their composites and that, in effect, there will be a different amount and kind of criterion deficiency in the estimates obtained from different raters, if not in different ratings made by a single rater. When limitations of the research study require the use of over-all ratings as the criterion, it would seem advisable to incorporate a careful definition of the important job elements in the directions for the execution of the ratings. This, if properly accomplished, should help to reduce the extent of criterion bias and to insure that the evaluations of the several raters are predicated on a more uniform constellation of elements than would otherwise be the case.

The foregoing comments appear to provide sufficient consideration of criterion deficiency in relation to ratings. Because of the effect of halo (discussed below), it is difficult to consider this problem intelligently. Ratings of different job elements are often found to be so highly interrelated that one suspects that the rater's impression of the ratee's competence is the only determining factor of general importance. Because of this effect, the authors do not wish to give the impression that adherence to the foregoing suggestions will produce substantial improvement in the results obtained.

An examination of research reports indicates that, in general, systematic job analysis is an initial step in the construction of job-sample criteria more often than in the construction of any other type of criterion measures. In spite of such systematic job analysis, it is the authors' opinion that important elements of on-the-job success are usually omitted from job-sample

criteria. Much of the difficulty in this respect arises because the job-sample criterion indicates how well the employee *can* perform under standard conditions rather than how well he *does* perform under normal work-a-day conditions. It could possibly be argued that for the validation of aptitude and achievement tests, as opposed to personality measures, this is precisely what is desired. Where on-the-job success is a function of personality variates, however, job-sample criteria are apt to be criterion deficient. As a result of the exclusive use of such criteria, truly valid measures of personality differences would be excluded from the battery selected for operating use. Thus, while the use of job-sample criteria may be recommended for the evaluation of production in certain types of situations, it is doubted that they should ever be used alone as a measure of over-all on-the-job success.

Criterion Contamination

Criterion construction based on arm-chair considerations of factors of availability, rather than on an analysis of the job situation, faces not only the danger of omitting important factors but also that of incorporating variables that are not measures of on-the-job success. While contaminants of the criterion occur in the process of deciding what to measure, it is in the construction of the actual scales, or other means of measurement, that the investigator most frequently faces the problem of contamination.

From our outline of the steps in criterion construction it will be noted that procedures and/or instruments for making such measurements must be devised as a second step following the determination of the job elements in need of measurements. In discussing contamination in relation to the major types of criterion measures, the broader meaning of the term as employed here should be borne in mind. The more conventional usage of the term limits it to contamination introduced by direct influence of predictor scores on the criterion. The basic example is the effect of knowledge of predictor scores on criterion ratings. Bellows (1) extended the meaning of the term to include such phenomena as opportunity bias and artificial restriction of production. In the present paper, as has previously

been noted, any source of variance in the criterion, other than error of measurement that is not a reflection of on-the-job success, is labelled "criterion contamination." Thus, our definition includes all extraneous elements in the criterion. However, several additional concepts will be introduced to aid in distinguishing between different types of contamination.

In production records, contamination most frequently occurs because factors beyond the control of individual workers considerably affect the amount of his production. This type of contamination has been referred to as *opportunity bias*.

Examples of opportunity bias may be cited for almost any type of job. In evaluating salesmen such bias may occur because of differences in the "goodness" of territory; in evaluating production line workers, it may occur because of differences between day and night shifts, in the location of the work site, in tools and machines, in the efficiency of supervisors, or in work-mates and repairmen. Differences between day- and night-shift workers may be substantial even though no differences exist as to potential productivity. If samples of production are obtained at different times for different workers, diurnal variations in productivity may bias the obtained criterion scores. Thus, it is known that work output definitely varies according to the time of day. Hence, records of production obtained on individuals at the time of optimal output would be biased in relation to those obtained at the time of minimal output. A comprehensive listing of the sources for opportunity bias is impossible. A careful analysis of the conditions of work of the various members of the experimental group during the collection of criterion measures is necessary to insure identification of such biasing factors.

The most important question to be answered with reference to opportunity bias is the degree to which it is test correlated or test free. First of all, the possibility that components of the experimental predictor battery were employed in determining who would be placed in the position where opportunity for high production record was greatest, should be checked. For example, if tests or other variables in the predictor battery were employed to determine which salesman obtained the best territory or which sales clerk was given the best counter, etc.,—as

might often occur if the selection procedures being validated were actually used in the operating selection program—the effect on validity of such biasing factors could be very considerable.

Suppose that 10 per cent of the variation in amount of production in a given job were due to some form of opportunity bias, and that placement in a position of greater opportunity had been in terms of a test employed for the prediction prior to the initiation of a research study. If, in this research study, this predictor was evaluated along with other experimentally constructed instruments, we can compute that the obtained validity (biased by the opportunity factor) would be .32, even though its actual validity were zero. It may be seen that the resulting contamination would be highly destructive to the objectives of the research study.

Even if no direct evidence of relationship is found between any predictor and opportunity bias in criterion scores, evidence of indirect relations should be sought. If seniority were to determine placement in the position of greatest opportunity, predictor variables such as age and experience would show heavily biased validity. Personal history items, bearing directly or indirectly on the length of experience or age, would have similarly biased validities. Other possibilities may be cited. Questionnaire items relating to marital status may appear to have high validity because a much higher percentage of non-married workers choose to work on the night shift. A measure of aggressiveness may falsely appear to be a valid predictor of sales records because the more aggressive salesman pushes himself into the advantageous sales-territories.

While the possibility that opportunity bias may be test-correlated should be thoroughly checked, it is probably generally true that the extent of the correlation will frequently be found to be negligible. Generally, in other words, opportunity bias will be test free and will attenuate or lower all validity coefficients but will not seriously distort their relative magnitude.

A second frequently mentioned contaminating factor in production records is the one introduced by limitations on rate of production. Such limitations may occur because of assembly line

production, because men work in teams, because of social pressure from other workers or from a number of similarly operating factors. These are not biasing in one sense of the term. If production of the faster workers cannot exceed that of the slower workers by more than 50 per cent, the observed difference is truly representative of the full advantage to be obtained by hiring the fastest in preference to the slowest worker for that given job situation. Of course, if the effect of a change in the composition of the efficiency of all members of the assembly line—or group—could be measured, the problem would be considerably changed. In order to obtain a measure of such effects it would be necessary to depart from the usual correlational methods of validating tests. It would be necessary to select groups with differing average productivity, to assign all members of each group to a given assembly line and to compare the mean productivity of these groups. In such comparison of groups, experimental controls would have to be established; that is, the conditions of work, and all factors influencing output, would need to be equalized for all groups, with variation between groups limited to the difference in predicted productivity. While the method for handling this special problem bears mention, extended discussion is not possible at this point.

While the effect of such factors is not contaminating in the sense indicated above, results due to the presence of such factors cannot, of course, be generalized to situations where such limitations are not present. The most obvious conclusion to be drawn when limitation on production is discovered, is that selection programs are likely to be of limited value.

To save time and money, a large proportion of the industrial selection researches are conducted on in-service personnel, i e., tests are administered to and criterion data are collected on personnel already in the employ of the sponsor. Such cross-sectional studies, while necessary, are always, to some degree, defective in experimental design. In practice, test scores are necessarily obtained prior to employment or to any other personnel action based on them. The conduct of cross-sectional studies may introduce two types of contamination when its results are applied to the employment situation. The first is a test contamination arising from the fact that both the on-the-

job experience, and the nature of the conditions under which the predictors are administered, may exercise considerable influence on test scores. This, being a predictor rather than criterion contamination, need not further concern us here.

The collection of criterion data on an in-service population may, however, introduce an experience-contamination which is of direct concern to us. Where the job is one in which production may be expected to rise with increased experience and there is considerable variability in the tenure of the validation population, the criterion will, of course, be contaminated with experience. If the predictors also include experience-correlated variables such as age, the contamination will be predictor correlated. If the tests are also experience-contaminated, such tests will show a spuriously high correlation with the criterion.

Validities of predictors such as information or proficiency tests, and knowledge of terminology, would tend to show positive bias in validity in cross-sectional validation studies. Knowledge of terminology and productivity would both tend to be greater in experienced than in inexperienced workers even though there might be no relation between the two measures among workers with equal experience. Bias of this nature may be avoided by testing prior to employment, by administering all tests to groups with constant amounts of experience, or by controlling experience statistically. The danger of such bias does not have bearing, obviously, on the utilization of experience prior to employment for the given job as a predictor.

Estimates of the reliability of production criteria are probably more often, and more seriously, distorted by biasing factors than are validities. Bellows (1) has pointed out that in many jobs where unequal opportunity seriously affects the production records of a category of workers, it is likely that a second measurement of the productivity of these workers will be obtained with the same biasing factors in operation and with the same workers showing spuriously high productivity. For example, if production records were obtained during two different intervals on a population of workers including those on day and night shifts, it would probably be found that day-shift workers produced more during both time intervals. The appar-

ent reliability of the production measure would be quite high even though its actual reliability were below usual standards of acceptability.

The construction of rating scales free of contamination presents, possibly, more serious problems than detection and elimination of contamination from production records. It should be stressed initially that all of the sources of bias discussed in connection with production records will probably also tend to influence ratings of productivity. It is possible, however, that raters may be successful in making allowance for some of these factors—opportunity biases, for example—and thus reduce their influence.

The most obvious and probably the most serious source of contamination peculiar to ratings arises because of the so-called halo effect.

The term "halo" implies that a spurious relationship between rated traits, attributed to a spread of the effect of the raters' attitude toward, or estimate of, the rater in one dimension over to his attitude toward or estimate of the rater in other, unrelated, dimensions. Various factors have been postulated as the source of the halo effect. Degree of personal liking is frequently mentioned as a possible source. Over-all impression, social prestige and outstanding achievement in a particular field are other possible sources. As yet there is no evidence allowing definite conclusions regarding the source of the halo effect. It may be regarded as established, however, that some factor or factors operate spuriously to increase the relationship between ratings on different characteristics.

Since the source of halo cannot be established, it cannot be regarded as necessarily a contaminating factor. Bingham (2), in discussing the role of halo in criterion ratings, expresses the belief that there are a number of situations in which the general impression that the individual makes on those he comes into contact with, can itself be an important criterion element. He concludes that halo should not, in all cases, be considered an undesirable attribute of criterion ratings.

It would be agreed by most, however, that Bingham's conclusion, even if correct, gives no sound solution to the problem of halo in criterion ratings. Even though halo reflects important

elements of on-the-job proficiency, it would be desirable to obtain adequate estimates of proficiency in the various aspects of the job, free of halo, in order to insure that separate job elements are properly weighted in arriving at an over-all composite.

Halo effect, if contaminating in nature, can become test correlated and thus assume considerable importance, particularly when the prediction battery includes ratings, personality measures *and* ability tests. In such a situation, the criterion and predictor ratings may show spuriously high correlation because of halo effect common to both. Personality measures may likewise show spuriously high validities through the prediction of the contaminating halo element. Since, at the same time, validities of ability-test scores would probably be attenuated, the partial regression weights for the entire battery would be considerably distorted. The tendency reported by Bingham and Freyd (3) for personality measures to show relatively higher validities against rating criteria and for objective tests to show relatively higher validities against production record criteria, may be explained, at least in part, by the biasing effect on the validities of personality measures noted above. It is particularly important to note that direct criterion contamination may result from a remote source. Variables which influence criterion ratings need not be members of the prediction battery in order to distort the validities and regression weights. If the variables which influence the criterion scores are correlated with any in the battery, the resultant criterion contamination will be test correlated; if such variables are uncorrelated with members of the predictor battery, the contamination will be predictor free.

A source of criterion contamination in ratings similar in its effect to opportunity bias arises from differences in the mean values obtained from different raters. Employees of a tough rater will receive lower criterion scores than will those of an easy rater. Normally, the resulting contamination will be test free. However, if assignment to various supervisors is made on the basis of test scores, such bias can be predictor correlated. Conrad (4) has contended that such differences in rater tendency are over-emphasized and that proper rating techniques will minimize such differences.

A basic source of contamination in ratings arises from the failure of raters or of rating-scale constructors to distinguish between those observations which constitute direct evidence of productivity and those which give only inferential evidence of productivity. To this source of criterion contamination the authors would like to give the name "*error of illation*." Thus, ratings on the efficiency of a carpenter based on observations on the skill with which he uses his hands, the air of assurance with which he handles tools or even the correctness of his choice of tools for each operation, are all inferential and without empirical evidence cannot be assumed to have high relationship to actual productivity. Even though such relationship were established, it could not be assumed that such trait ratings could be substituted for direct measures of productivity without biasing effect on the validation results.

In designing forms that incorporate scales for the measurement of such traits as manual skill, industriousness and ambition, the psychologist promotes this form of contamination. Ratings on such traits give rise to the danger that resulting evaluations may not only have been inferred, but that they may have been inferred, in large part, from events observed in a social situation or in other situations having no necessary relation to on-the-job productivity. Evidence on the highly specific nature of psychological traits from studies by Hartshorne and May (10) are pertinent in showing the dangers of such bias.

The tendency of rater to consider the symptoms of productivity rather than productivity itself can probably never be entirely eliminated. It should be possible in many work situations, however, to identify the individuals with the greatest opportunity to observe the actual production element and to orient the scales so that the evaluations given by the rater involve as few deductions and as much direct observation as possible. The directions and content of the scales can be so oriented that they specifically request the rater to base his evaluation on direct observation of results. Even though it is improbable that the desired purpose will be entirely accomplished, the technician should at least not be guilty of encouraging a tendency toward inference rather than direct observation.

by phrasing his directions and scales in terms of indirect or inferential content.

Having constructed scales oriented toward direct evidence of productivity and having determined who is in the best position to evaluate each directly, the technician may take one additional step to help reduce errors of illation. The raters may be instructed, well in advance of the collection of the criterion ratings to observe and to take note of behaviors falling in the areas to be rated. Mention should be made of the fact that such oriented observation is an integral part of the "critical incident" technique mentioned above.

The bias introduced by the illation error is probably very often test correlated. Trait ratings obtained prior to employment might give excellent prediction of ratings of traits *thought* desirable for efficient performance but not actually related to quantity and quality of production. If so, nothing will have been demonstrated. Personality tests related to the traits thought desirable would similarly yield inflated validity coefficients.

The danger of contamination in use of achievement-test scores, as criteria of success in training or in school, are considerable. Probably, also, such contamination will be test correlated. Frequently, information tests are employed along with aptitude and other measures to predict achievement in training. Such achievement is also measured by an information test administered at the end of training. Test constructors working on both the predictor- and criterion-information tests may well employ the same source material for constructing items and may well both err in the same direction in selecting items irrelevant to or unimportant in the actual training process. Such common but irrelevant content in the predictors and criterion can naturally be expected to produce test-correlated contamination.

It is probable that a similar biasing effect is often obtained in relating any ability-test measures to success in training. Generally speaking, ability-test scores have shown uniformly high validities in this area. Such validities are suspect, however, since they are obtained by relating initial test scores to measures of proficiency after training. Woodrow (17) has shown that initial test scores show little relation to improvement with practice. He has also shown that general-intelligence-test scores

(often interpreted as measures of learning ability) have little if any relation to *improvement* in scholastic achievement. Since the essential problem is the prediction of benefit derived from training, lack of evidence contradictory to that reported by Woodrow suggests that predictors of training success or training improvement have doubtful validity for that purpose. The selection instruments may, of course, still have value in predicting on-the-job success. To assume such validity, knowing only that the predictors correlate with estimates of achievement in training, assumes that achievement in training is highly related to on-the-job success. Little, if any, research has been reported demonstrating a positive relationship between training success and later success on-the-job. The low correlation of the academic achievement of West Point Cadets (8) with later success as Army officers, argues strongly that training success cannot be assumed to have appreciable relationship to success on-the-job.

Job-sample criteria are possibly less subject to contamination than any of the criteria discussed. Opportunity can be carefully controlled. Halo, effect of easy-hard raters, etc., can be reduced to a minimum. Ratings of work products, while subjective, differ in character from ratings of individuals. In rating work products, raters need not know the individuals whose products are being evaluated and the effect of personal likes and dislikes of the rater can thus be eliminated.

However, because of the similarity between the test-like character of the situation under which the job-sample measures are obtained, and the usual conditions under which tests in a predictor battery are administered, contamination, test-correlated in nature, is probably often present in job-sample criteria. Individuals who become overexcited or nervous in the one situation may tend to show the same type of behavior in the second. Similarly, individuals who put forth greater effort when being watched, would be apt to do so in the type of situation in which both tests are administered and job-sample measures are obtained. It is possible, also, that if tests and job-sample performances are obtained on the same day, factors peculiar to the day of testing will act as test-correlated contamination and introduce a positive bias into the validity coefficients.

A type of criterion scale in which the possibility of contamina-

tion is easily overlooked is that in which "high" and "low" groups are employed. For some not-too-apparent reason, investigators seem to feel that by selecting extreme groups they have circumvented the problem of contamination and have secured "pure" cases. It is strongly emphasized that the selection of such groups is based on a continuum either actual or implicit. Extremes on this continuum may be extreme on a contaminating element as well as on the "true score" component of this continuum. Such measures should be as carefully scrutinized for contamination as any continuous criterion.

Investigators often show a similar tendency to neglect problems of bias in using a group-membership criterion. In the situation in which members of one occupation are compared with members of other occupational groups, or with the general population, the opportunity for criterion contamination is extensive. Where the "in" group has been test selected and the same or correlated predictors are included in the experimental battery, presence of extensive test-correlated contamination is almost certain.

Even where prejudices, rather than tests, dictated entrance into the occupational group, predictor-correlated contamination may be expected. If an executive, for example, arbitrarily ruled that all messengers coming into the firm should be high-school graduates, and employed messengers were compared with some general group, education and educational achievement tests would show substantial validity even though their true validity were negligible.

The composition of occupational groups is determined by factors determining the initial choice of occupation and by attrition after such initial choice. Factors responsible for choice of occupation are almost certainly a source of contamination; those responsible for attrition *may* be of value for criterion purposes. It is not usually possible to obtain any reasonably exact information concerning the major factors in either case. Because of this lack of information, if for no other reason, such a criterion is suspect.

The preceding discussion of contamination could not be completely comprehensive. In any individual research study, contamination peculiar to that study may be discovered. We

have endeavored, however, to clarify and illustrate the nature and effect of the more important general factors.

Criterion Scale Unit Bias

While the presence of scale-unit bias in criteria has frequently been recognized, particularly in connection with ratings, the general problem has not been extensively explored. A review of the psychological literature provides little evidence allowing an estimate of the prevalence or seriousness of scale-unit bias in the criteria of validation studies.

Basically, it is believed that the problem centers in the absence of an adequate rationale. There is no generally accepted means of judging the presence or absence of scale-unit bias available to the investigator desirous of evaluating the relative merits of various possible types of scale units or scaling procedures.

Possibly the only widely used standard of adequacy of scale units is the degree of approximation of the obtained frequency distribution to a normal curve. While this standard may be of some value in avoiding serious distortion of scale units, it must be remembered that normality is always an assumption. Standards are needed that will allow checking the adequacy of scale units in a particular example without the necessity of such an assumption. From a logical view point, a standard for judging presence or absence of scale-unit bias that applies to the shape of the frequency distribution is in any event defective. The distribution form is a function of the population involved as well as of the scale units. Normality should certainly not be considered desirable where there is strong presumptive evidence that selection of cases has occurred.

It is fortunate that, in general, product-moment validity coefficients do not seem to be seriously affected by alteration of scale units so long as rank order is unchanged. When test scores are converted to normalized form, or when ratings obtained in rank-order form are normalized, product-moment validities are usually very little altered.

While the product-moment validity for the entire range is probably little affected by scale-unit distortion validity, indexes computed for particular points of cut on the predictor may

be seriously affected. Where scale-unit bias is suspected, such coefficients should be interpreted with caution.

We might note also that a heavily skewed criterion distribution, if established as genuine, would have implications of some significance for efficient selection. Individuals on the tail of a skewed distribution could undoubtedly be identified with greater confidence than those in the same percentile point on a normal curve. Thus, while the problem of scale units may not be of great significance when conventional methods of analysis are employed, a solution to the problem that would allow identification of highly skewed distributions with confidence could lead to improved efficiency of selection through different methods of analysis.

From the criterion point of view, the scale-unit problem reduces to one of establishing units which represent equal increments in terms of the over-all efficiency of the organization. This point will be elaborated by the authors in a forthcoming paper on that topic.

In terms of the efficiency of the organization, production records appear to be relatively free from scale-unit bias. An additional object produced has equal value whether it increases the productivity measure of an individual from 1 to 2 or from 99 to 100. A given error is just as costly no matter whether it increases the error score from 4 to 5 or from 19 to 20. Such units have meaning in their own right. Even in the evaluation of quality of production, differences in quality can be assigned values having direct meaning if the resulting objects of differing quality are eventually sold for different prices. Quality differences would then acquire a quantitative monetary value. This cannot, however, always be accomplished.

Ratings are subject to a number of forms of criterion scale-unit bias. Piling at the upper end of the scale, failure to employ the lower scale units, piling in the center of the scale and other defects have all been frequently reported in the literature. Since these tendencies appear in wide varieties of rating situations, it seems reasonably certain that they are distortions of the scale units and are not due to the nature of the true distribution of the degree of productivity in the job element being rated.

Lack of information as to the true or proper distribution

form considerably hampers the solution to the problem of scale-unit bias in ratings. While it seems reasonably certain that the scale-unit biases mentioned above do often occur, it is difficult to judge in any particular instance when a rating scale is free of scale-unit bias and, more particularly, the nature of such bias as may be present.

In the absence of evidence to the contrary, a normal distribution of criterion rating scales would usually indicate freedom from scale-unit bias. If the distribution of production records, on the job element being rated, is known from other research studies, such distributions would probably provide a sounder basis for judging the adequacy of the distribution form of the criterion ratings than would the normal curve.

In the use of order-of-merit rankings it is apparent that the form of the distribution is forced and that equal numbers of individuals fall within each interval of a given magnitude. If, however, rankings are obtained from a number of different raters it will usually be found that the average of the rankings will approximate the normal curve to a satisfactory degree.

No problems of scale-unit bias arise which are peculiar to job-sample criteria. If job-sample criteria are scorable in production units, comments made with reference to scale-unit bias in production units will apply here also. If scoring is subjective, problems similar to those encountered in rating scales will occur. It seems probable, however, that scale-unit bias will be less extreme than that occurring in direct evaluation of individuals. The direct evaluation of production has the added advantage that in some cases it can be divorced from the individual to some degree and thus escape, in part at least, some of the biases which stem from the interpersonal relations between rater and ratee.

Achievement tests employed as criteria involve scale-unit biases of a nature peculiar to continuous variables obtained by summing a number of dichotomous items. Where rating scales are so constructed they will also be subject to this form of scale-unit bias. Variation in the difficulty level (i.e., percentage of raters checking a given item) will have considerable effect upon the distribution form of the total score. The effect here is very similar to the effect of item-difficulty distribution on factor

structure of tests discussed in some detail by Ferguson (6) and Wherry and Gaylord (16). The frequency of occurrence in the population of a component element of a criterion scale is analogous, in other words, to the difficulty level of component items of a test insofar as the statistics of their interrelations are concerned. If a criterion consists of high difficulty elements, it will tend to correlate more highly with tests also consisting of high difficulty items and less highly with tests consisting of low difficulty items.

When a criterion variable consists, then, of a number of discrete items, the investigator should take care to insure that the difficulty level or the "frequency of occurrence" level corresponds to the frequency of occurrence of the job element in the work situation. To accomplish this purpose, a difficulty distribution should probably be determined for each set of criterion components, and the number of observations or measures at each level should be made to adhere to this predetermined distribution.

Criterion Distortion

An additional source of bias, which we have referred to as "criterion distortion," arises as a result of the improper assignment of weights to the several elements. More broadly defined, criterion distortion would include all of the other types of bias discussed. Thus, *criterion deficiency* is the assignment of weights of zero to elements that should in reality have non-zero weights. *Criterion contamination* is the opposite error; the assignment of non-zero weights to elements that merit no consideration. *Criterion scale unit bias* in effect assigns different weights to different parts of the continuum of the given criterion element.

A number of techniques have been proposed for determining the proper weights for criterion elements. We may do well to examine several of the procedures that have been proposed and to investigate the type of situation in which each is most appropriate.

Horst (11) and Edgerton and Kolbe (5) have proposed procedures which, in effect, operate to maximize the reliability of the over-all criterion. The assumption implicit in these techniques is that all criterion elements measure the same basic variable and that the lack of perfect correlation between them

is attributable to error of measurement. This procedure would thus be quite applicable to situations in which the criterion consisted of several measures of the same attribute, such as ratings by different observers of the same trait. It seems evident, however, that this technique should never be employed in combining elements which attempt to assay behavior on different continua. Unfortunately, the technique has often been used for this latter purpose. It is the author's opinion that the chief advantages of employing techniques developed by mathematical derivation lie in the thorough and explicit manner in which the assumptions must be stated. If the assumptions used are ignored in applying the technique or formula developed, the mathematical development is, in a sense, disadvantageous in that it lends prestige to a formula completely unsuited to the particular application.

Where no objective basis exists for the establishment of the relative weights of criterion elements, weights obtained by Toops' (14) method of guessed Beta weights is, in the authors' opinion, superior to an unweighted raw or standard score sum. Toops proposes averaged estimates of the judged importance of the various criterion elements as a means of weighting, the judges being those personnel in the sponsoring agency having the best knowledge of the implications of various criterion elements for the efficiency of the organization as a whole. There are a number of technical problems involved in making clear to the judges the proper basis for guessed Betas. Consider, for example, the problem of obtaining weights for combining the number of production units and the number of errors. Should the evaluations requested be phrased so that raw-score weights are obtained or so that standard-score weights are obtained? Since the judges will probably not understand the effect of differences in the standard deviation of criterion elements on their effective weighting, how can bias from this source be avoided? In spite of these problems in technique the method provides a direct approach to the basic problem of weighting criterion elements. In addition, its sponsor acceptability should be high. These factors, in the author's opinion, suggest the advisability of a more extensive use of this technique.

It should be stressed that the common practice of computing

separate validity coefficients for the various subcriteria is equivalent, in the final analysis, to a method of combining criteria scores. It differs in that the experimenter avoids a formal procedure. Instead, he merely looks at the validities against the several criteria and decides on the tests which are to constitute the selection battery. Such a procedure has the effect of concealing from the research worker himself the fact that he is deciding the relative importance of the sub-criterion variables. Usually, the investigator will decide to include several tests for the prediction of each of the criteria, and will fail to consider the relative importance of the criteria or to evaluate properly the effect of the intercorrelations and validities or the partial regressions for predicting a composite. The problem is thus evaded rather than solved. In general, a formal solution will at least make explicit the basis for the decisions concerning the relative importance of the several criteria and will avoid incidental errors which may creep in because of carelessness in the subjective handling of the data.

A suggestion by Otis (12) may, in particular instances, lead to a more meaningful combination of subcriterion scores than would result from the application of any of the procedures so far mentioned. Otis pointed out that, in key-punch operation, it was discovered that the correction of an error required the time equivalent to that needed for punching 14 cards. He suggested, consequently, that a total over-all production index could readily be obtained simply by subtracting 14 cards for every error made.

The method of combining criteria suggested in this particular instance, is not exactly a technique and does not suggest any uniform procedure that can be widely employed. It does suggest, however, that detailed examination of the relationship between the different work units measured and the organization of the over-all productive process will often suggest that certain different sub-criteria are, or can be, expressed in units which are equivalent in their effect upon the total productivity of the organization.

The effect of the use of inappropriate weights for criterion elements, as with other forms of bias, will depend upon the extent to which it is predictor free or predictor correlated. The

overweighting of any given element will naturally afford undue weight to the predictors that have the highest correlation with the overweighted element or elements. Conversely, the predictors that correlate highest with underweighted elements would be given inadequate weight. Prediction would hence be distorted and while in a selection problem, for example, the predictors would align the population in accord with the criterion as weighted, this alignment would be at variance without the "true" criterion.

The reader may readily judge that the authors consider most procedures for criterion combination in current use to be not wholly adequate. This appears to be an area particularly in need of further research.

Summary

This paper proposes a classification of criterion bias into four main categories:

1. Criterion deficiency
2. Criterion contamination
3. Criterion scale unit bias
4. Criterion distortion

Each category is discussed in terms of the steps in the criterion-construction process in which it is most likely to occur. Each is also briefly related to the several kinds of criterion measures. Each type of bias is also considered in relation to various types of criterion measures. The importance of distinguishing between bias that is test free and bias that is test correlated is emphasized. In discussing possible biasing factors, the test-free or test-correlated character of the biasing factor has received continual emphasis.

Biasing factors reported in the literature have been considered. Additional concepts have been advanced by the authors.

REFERENCES

1. Bellows, R. M. "Procedures for Evaluating Vocational Criteria." *Journal of Applied Psychology*, XXV (1941), 499-513.
2. Bingham, W. V. "Halo, Invalid and Valid." *Journal of Applied Psychology*, XXIII (1939), 221-228.
3. Bingham, W. V. and Freyd, M. *Procedures in Employment Psychology*. New York: Shaw, 1926.

4. Conrad, H. S. "The Personal Equation in Ratings: A Systematic Evaluation." *Journal of Educational Psychology*, XXIV (1933), 39-46.
5. Edgerton, H. A. and Kolhe, L. E. "The Method of Minimum Variation for the Combination of Criteria." *Psychometrika*, I (1936), 183-187.
6. Fergeson, G. A. "The Factorial Interpretation of Test Difficulty." *Psychometrika*, VI (1941), 67-77.
7. Flanagan, J. C. "Critical Requirements: A New Approach to Employee Evaluation." *Personnel Psychology*, II (1949), 419-425.
8. Gaylord, R. H. and Russell, E. "West Point Evaluative Measures in the Prediction of Officer Efficiency," in preparation.
9. Guilford, J. P. "New Standards for Test Evaluation." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, VI (1946), 427-438.
10. Hartshorne, H. and May, M. A. *Studies in Deceit*. New York: Macmillan Co., 1928. Page 414.
11. Horst, P. "Obtaining a Composite Measure from a Number of Different Measures of the Same Attribute." *Psychometrika*, I (1936), 53-60.
12. Stead, W. H., Shortle, C. L., et al. *Occupational Counseling Techniques*. New York: American Book Co., 1940.
13. Stuit, D. B. (Ed.) *Personnel Research and Test Development in the Bureau of Naval Personnel*. Princeton: Princeton University Press, 1947.
14. Toops, H. A. "The Selection of Graduate Assistants." *The Personnel Journal*, VI (1928), 457-472.
15. Toops, H. A. "The Criterion." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, IV (1944), 271-297.
16. Wherry, R. J. and Gaylord, R. H. "Factor Pattern of Test Items and Tests as a Function of the Correlation Coefficient: Content, Difficulty, and Constant Error Factors." *Psychometrika*, IX (1944), 237-244.
17. Woodrow, H. "Interrelations of Measures of Learning." *Journal of Psychology*, X (1940), 49-73.

AN INVESTIGATION OF TWO HYPOTHESES REGARD- ING THE NATURE OF THE SPATIAL-RELATIONS AND VISUALIZATION FACTORS¹

WILLIAM B MICHAEL

Princeton University

and

WAYNE S ZIMMERMAN and J. P GUILFORD

University of Southern California

PRIMARILY as a consequence of the factorial analyses of tests of intellectual abilities, the construct of a spatial and/or visual ability amenable to psychological measurement has received increasing attention in recent years. During the past twenty-one years, at least a score of investigators have identified in their writings a space factor. In a pioneer study, Thurstone (16) included among his seven primary mental abilities a factor labelled S, which he characterized as a "facility in spatial and visual imagery,"—a factor which he likened to the spatial or visual group factor found by Kelley (13) in earlier experiments. The same factor was identified in other studies carried out subsequently by Thurstone (17) and by Thurstone and Thurstone (18).

During World War II members of the psychological research units of the Army Air Forces devoted a considerable amount of time and effort to the development of tests of the "spatial-visual" type to be used in the selection of men for air-crew positions. Several factorial studies which have been described in a research report of the AAF Aviation Psychology Research Program, edited by Guilford (3), have indicated that the vari-

¹ The first-mentioned author wishes to express his sincere appreciation to the Social Science Research Council for kindly making available . . . completion of this investigation. The authors are indebted to . . . who generously granted permission to have several of his tes . . . they might be included within the battery. Grateful acknow . . . staff of the Department of Psychology at Rutgers University for their interest in the study, their cooperation in making subjects available, and their assistance in administering a number of the tests. To all Rutgers students who participated, special thanks are extended.

ance associated with Thurstone's spatial-visualization factor may be separated into two apparently independent factors identified to be spatial relations and visualization (visual manipulation). In fact, in addition to these two factors (abbreviated by the symbols S_1 and V_z), two other less definite space factors, S_2 and S_3 , and a factor tentatively identified as visual memory, also appeared in several analyses.

In two recent studies both Fruchter (2) and Dudek (1) have found separate factors of spatial relations and visualization. In his investigation as to the nature of verbal fluency Fruchter reanalyzed a sub-matrix of twenty tests selected from the battery of fifty-seven variables employed by Thurstone in his classical study previously cited (16). He found two independent factors which he described as being spatial-relations and visualization.

Referring to the same Thurstone study, Zimmerman (3) pointed out that further rotations of the residual axis (Number XII) with other axes which defined meaningful factors would produce a promising factor of visualization. Just recently, Zimmerman (in his unpublished doctoral dissertation) has rerotated the twelve centroid axes for all fifty-seven variables and has confirmed his initial belief that both a spatial-relations and a visualization factor would appear.

Problem

THE purpose of the investigation was to test the validity of two (apparently unrelated) hypotheses that purport to represent differences in the psychological properties of the factors of spatial-relations and visualization as reflected by corresponding differences, both in the respective contents of two types of tasks and in the respective work procedures required of the subjects for successful completion of them. Each type of task consisted of a group of three tests. Within each of the two groups of tests employed in the study there appeared to be not only a similarity in the format of the test items, but also a common approach or operation demanded of the examinee.

In broad outline the plan followed in the investigation was to incorporate within a test battery two groups of tests which the investigators believed to be representative of the psycho-

logical operations involved in the hypothetical statements as to the nature of the spatial-relations factor and of the visualization factor. In the selection of each test to be incorporated within a group, introspection was freely employed as an aid to the determination of the psychological processes used in the subjects' performance upon a test—the same processes supposedly as those indicated in the relevant hypotheses. To these six tests (i.e., two groups, each consisting of three tests) were added eight reference tests of fairly well-known factorial content to aid in the identification of those portions of variance in the six tests that were associated with other factors such as verbality, numerical facility, reasoning, and perceptual speed. The inclusion of other factor tests served not only to identify what probably without their presence would be large amounts of specific variance within each of the six tests, but also to indicate the relative degree of purity of each of these six tests with respect to the function it was hypothesized to measure.²

A sufficient, though not necessary, condition for the tenability of each of the hypotheses, would be that in the factor-analysis procedure each of the two groups of three tests would define a factor. Moreover, this factor should not appear to be weighted in other tests of the battery that were selected to measure other factors. If one or more tests within either group should be weighted substantially in variance associated with another factor, the evidence for the corresponding hypothesis would be less clear-cut, but not necessarily lacking. It would be quite possible, if not almost certain, that one or more of the three tests within a given group might be factorially complex. At the same time, however, all three tests within a given group might contain substantial amounts of variance in one factor that did not appear in any of the other eleven tests.

Hypotheses

The factor of spatial relations was hypothesized to represent the ability to comprehend the *arrangement* of elements within

² It was also thought to be very desirable to determine whether tests of the type used by the AAF and Thurstone's tests held in common factors identified as being the same. This is the first study the writers know of that will serve to check upon the belief that many of the Thurstone primary abilities and the AAF factors are identical. Only the Thurstone space factor is here called into question.

a visual stimulus pattern, primarily with reference to the human body. Thus, an important implication in the ability to perceive spatial arrangements is that the subject is able to distinguish whether one object is higher or lower, left or right, or nearer or farther than another within the same field. Through the presentation of two simulated views of a stimulus pattern, a test item may be constructed such that there is a systematic relationship between the order of elements within the first spatial pattern (the stimulus component of a test item) and the order of elements within the second pattern (the response component of a test item).

For example, in Thurstone's *Cubes* test the examinee is asked to recognize whether the designs on the sides of a second cube can hold the same relationship to one another as they do on the first cube. By noticing within each cube the left-right, top-bottom, and front-back interrelationships of the faces, the subject is able in each item to refer the locations of three designs on three exposed faces of one cube to the locations of designs on the faces of the other cube. In Thurstone's *Flags* test the examinee is required to tell whether the exposed faces of two American flags of identical size can represent the same side of the flag. Relating corresponding left-right and top-bottom boundaries (outlines) of the two flags appears to be an important aspect of the solution. Similarly, in Guilford and Zimmerman's test of *Spatial Orientation* a premium is placed upon the examinee's maintaining the correct relationship of objects to one another in background scenery that has been viewed twice from a motorboat—first before and then after its prow has moved up or down and/or left or right. In the test the examinee is asked to determine the relative amount and direction of movement of the boat corresponding to changes in the two views of the background setting.

The factor of visualization was hypothesized to represent an ability that requires the mental *manipulation* of visual images. In contrast to another factor identified as visual memory (3), which appears to be a static or reproductive form of visualization, the factor referred to as visual manipulation, or simply visualization, is dynamic. This visual manipulative ability appears to be present in the solution of problems in which the in-

dividual finds it necessary mentally to move, rotate, turn, twist, or invert one or more objects. Following the performance of the presented manipulation the individual is required to recognize the new position, location, or changed appearance of the object or objects.

Three tests selected to yield evidence for the second hypothesis included two by Thurstone, *Punched Holes* and *Form Board*, and one by Guilford and Zimmerman, *Spatial Visualization*. In the test of *Punched Holes* the examinee is presented a symbolic representation of a folded sheet of paper into which one or more holes have been punched and is required to imagine where the holes will be when the sheet is unfolded. In the second Thurstone test the examinee apparently finds it necessary in each item mentally to turn, rotate, or invert two or more flat geometric figures in such a way that they can be placed together to fit within the outline of a larger geometric figure. In each of the tests, the examinee is asked to record the final positions respectively of the holes and of the geometric figures. In the test of *Spatial Visualization* the subject is required mentally to turn, tilt, or rotate a three-dimensional object—an alarm clock—drawn on a sheet of paper into a final position according to written instructions. As alternative responses the pictures of the clock are presented in five positions, one of which is correct. (A more detailed description of these three tests follows in the next section.)

Whereas in the two Thurstone tests the examinee is required to draw in his solution to the problem, in the third test he merely selects as his solution one of five choices presented. It is quite likely that in addition to measuring visual manipulative ability other factors are involved in the three tests—factors reflecting the manner in which responses to the items are recorded.

Another important difference in the nature of the psychological processes hypothesized for the spatial relations and visualization factors was that of speed of response. As indicated by findings in the AAF Aviation Psychology Program, the tests thought to measure the spatial relations factor were administered with fairly short time limits, but those tests thought to measure visualization were given with fairly liberal time al-

lowances. The spatial relations factor was considered to demand a fairly rapid decision on the part of the examinee as to the spatial position of objects with reference to his own location; whereas, the visualization factor was believed to be represented in problems requiring a more deliberate and less automatic approach. In part, such a distinction may be a function of the complexity of a task (i.e., the number of steps entering into the performance of an item), the more complex tasks requiring visualization for their solution.

Concerning the psychological properties of spatial-relations and visualization factors, one other important difference has been suggested in the work of one of the psychological research units of the AAF, as follows:

The idea for Flight Orientation [a test] was proposed at the time Aerial Orientation [another test] was being developed. It was hypothesized (1) that the ability visually to maneuver an airplane as if from a position outside the cockpit is a manipulatory-visualization ability and (2) that the ability to imagine maneuvers taking place as if the examinee were within the cockpit is a spatial-orientation ability.

The Aerial Orientation test utilized cockpit views of outside terrain to be matched with depicted plane attitudes; the visualization-of-maneuvers tests involved only views of airplanes seen from a position outside of the cockpit. . . . Flight Orientation was designed to fulfill the requirements of the indicated variation—a test that would utilize only cockpit views of outside terrain. From hypotheses given above, it follows that Aerial Orientation should measure a combination of manipulatory-visualization and spatial-orientation abilities, while Flight Orientation should be a purer measure of the ability to orient in space (3).

That the two groups of tests selected for investigating the validity of the hypotheses may actually contain variance in both the spatial-relations and visualization factors would not be surprising, inasmuch as many subjects on the basis of their own introspective reports revealed that they made use of the two psychological processes associated with the respective hypotheses in tests selected to represent the implications of only one hypothesis. For example, if in the *Flags* test the subject is able, so to speak, to pick up the flag, move it, turn it about as if he actually has a model in his hands, then visualization is believed to be dominant. On the other hand, if the subject is

concerned primarily with the left-right and top-bottom orientation of edges of flags with respect to his own position, or if he has to move himself to a different position as in cocking his head to one side, then a spatial factor is believed to be more prominent.

Similarly, in the *Cubes* test if the subject reports he picks up the first cube and rotates it into a final position which matches (or cannot match) the second cube, then the visualization process is dominant. However, if he attempts primarily to interrelate the positions of the sides of the cubes with respect to his own position, or if he appears to project himself amidst the cubes as if he were walking about them and relating the locations of various sides with respect to his own position, then the spatial-relations factor is probably operative. It may well be that in the spatial-relations factor empathy plays an important role in the relating of the position of objects to one's own location, whereas in visualization the individual obtains first from a distance an overall view of the objects to be manipulated and then employs perhaps some rather restricted kinesthetic imagery in the imagined use of hands for moving the objects into their required positions.

Despite the apparent differences in approach employed by many subjects, it did appear that the two groups of tests chosen represented reasonably well a distinction between the psychological processes hypothesized. If a test did involve to a substantial degree the use of two or more psychological abilities, it was thought that the factor-analysis procedure would reveal such a fact.

Tests

In Table 1 are presented the names of the fourteen pencil-and-paper tests employed in the battery, the maximum number of items that could be attempted, the plan followed with respect to "speed" or "power" time-limit, the actual working time allowed, and the scoring formula used. The numbering of the tests in the tables, as well as in the following description of content and procedure, corresponds to the order of administration. During the first, second, third, and fourth testing periods, respectively, the following groups of tests were ad-

ministered: 1, 2, 3, and 4; 5, 6, 7, 8, and 9; 10 and 11; 12, 13, and 14. An ample number of practice exercises preceded the main body of each test. Further information concerning several of the tests may be found both in a manual (11) and in the literature (12, 16, 18, 20). It is believed, however, that the descriptions given will suffice for the interpretation of the factors to be presented.

1. *Guilford-Zimmerman Verbal Comprehension*.—This is a vocabulary test in which the examinee is required in each item to

TABLE 1
The Test Battery: Descriptive Data

Name of Test	Number of Items	Timing Plan (Speed or Power)	Working Time	Scoring Formula
1. Guilford-Zimmerman Verbal Comprehension	30	Power	10 min.	R-W/4
2. Guilford-Zimmerman General Reasoning	14	Power	13 min.	R-W/4
3. Guilford-Zimmerman Numerical Operations	127	Speed	5 min.	R-W
4. Guilford-Zimmerman Perceptual Speed	48	Speed	3 min., 45 sec.	R-W
5. Guilford-Zimmerman Spatial Orientation	60	Speed	8 min.	R-W/4
6. Thurstone [Verbal] Completion	30	Power	7 min.	R-W/4
7. Thurstone Number Series	20	Power	8 min.	R
8. Thurstone Identical Forms	40	Speed	3 min., 15 sec.	R-W
9. Thurstone Cubes	42	Speed	5 min.	R-W
10. Thurstone Flags	48	Speed	4 min.	R-W
11. Guilford-Zimmerman Spatial Visualization	40	Power (limited)	15 min.	R-W/4
12. Thurstone Punched Holes	10	Power	7 min.	R
13. Thurstone Pattern Analogies	20	Power	10 min.	R-W/4
14. Thurstone Form Board	28	Power	7 min.	R

choose among five words, all matched with respect to difficulty, the one word which most closely approximates the meaning of the stimulus word. Items increase in difficulty progressively from the beginning to the end. Even numbered items were omitted. Responses were recorded on a separate answer sheet. Most examinees attempted all items.

2. *Guilford-Zimmerman General Reasoning*.—This test is composed of arithmetical-reasoning problems similar to those encountered in courses in general mathematics, elementary algebra, and intermediate algebra. Diagrams accompany a few of

the problems. Numerical work is kept to a minimum. Five multiple-choice responses are presented with each problem statement. Items increase in difficulty level progressively from the beginning to the end. Even-numbered items were omitted. Responses were recorded on a separate answer sheet. Most examinees attempted all assigned items.

3 *Guilford-Zimmerman Numerical Operations*—This test is in four parts, consisting of numerous simple problems (of about the same difficulty level) involving respectively the four fundamental operations of addition, subtraction, multiplication and division. Emphasis is placed in the directions upon the need for both accuracy and speed of work. Subjects were told to begin with the part upon addition, to work every item, and to go as far as possible in the allotted time. Only a few subjects reached the fourth section upon division. Responses to the items were printed in spaces on the test booklet adjacent to the problems.

4 *Guilford-Zimmerman Perceptual Speed*.—This test requires the examinee to match a visual object of a familiar shape and of detailed design with one of five other visual objects of a common category (e.g., automobiles, boats, hats, shoes). Four of the five response objects resemble rather closely the stimulus object, but differ from it in certain minor details of shape and/or design. For each common category two parallel sets of visual objects—four stimulus and five response objects—are arranged in two parallel columns. To each one of the four stimulus figures in the first column corresponds one of the five response figures. Thus, four responses are scored for each item of homogeneous content. All items represent a low level of difficulty. Answers to the items were marked on the test booklet in spaces adjacent to each stimulus object. The examinees were told to go as far as possible in the allotted time. No examinee finished.

5 *Guilford-Zimmerman Spatial Orientation*.—This test requires an examinee to determine how the position of a boat has changed in a second picture from its initial position in a first picture. In each picture the prow of the motorboat, in which the examinee is told to pretend to be riding, is shown along with background scenery consisting of water, or a silhouetted shore line, and in some instances of other boats intervening be-

tween the shore line and the prow of the motorboat, which is in the extreme foreground of the picture. In the sample problems described in detail in the directions, the position of the prow in the second picture, with respect to the spot of background sighted over it in the first picture, is taken as the primary reference guide for determination of the direction and amount of subsequent up-down and/or left-right motion of the boat. Movement is also indicated by accompanying shifts in the location of elements within the pattern of visible background scenery. The boat is actually stationary with respect to any forward-backward motion. To each set of two pictures five alternative responses are presented. Each response is represented by (1) a dot designating the aiming point, the initial spot in the background sighted right over the point of the prow in the first picture, and (2) an arc (of about 45°) representing the location of the prow in the second picture with reference to the aiming point. One of the five responses shows the correct change in position of the prow of the boat with respect to the aiming point. All examinees were instructed in the limited time allowed to attempt as many items as possible. As in all other speed tests, answers were recorded in the test booklet. The difficulty of the items tends to increase for items further removed from the beginning of the test. No one attempted every item.

6. *Thurstone [Verbal] Completion*.—This test is one adapted from the *Psychological Examination of the American Council on Education*. Representing, probably, a combination of verbal comprehension and verbal fluency, it presents for each item the definition of a word, the number of letters in the word, and five alternative letters (responses), one of which represents the initial letter of the defined word. Although the items differ considerably with respect to difficulty, most of the defined words are familiar to college students. Responses were recorded on the page of test items. Nearly every subject attempted all items.

7. *Thurstone Number Series*.—Found to be loaded in a factor identified by Thurstone as induction, this test requires the subject to determine a rule for each item. Numbers are presented in a row with two blanks inserted. The task is to find the mathematical principle by which the number series

is formed and to insert in the blank that number which is appropriate. The difficulty level of items increases in relation to the position of the item from the beginning of the test. Responses were recorded on the test sheets in the blanks inserted at various positions within the different number series. Most of the subjects attempted all items. One point of credit was given to each blank correctly filled (two points per item being maximum score).

8. *Thurstone Identical Forms*.—This test resembles rather closely the fourth test, *Perceptual Speed*, in that the examinee selects from a row of five similar appearing figures that one which is exactly the same as the stimulus figure. Slight differences in color design and in shape appear among the five response figures. In this test the items are also homogeneous with respect to difficulty. The number corresponding to the sequential position of the response selected was recorded on the test page in a box to the right of the row of response objects. Only a few examinees reached the last few items.

9. *Thurstone Cubes*.—In this difficult test the subject is asked whether two drawings can represent the same cube on each face of which there is supposed to be a different design. In each of the two drawings the designs of three faces of the cubes are always exposed. If the two drawings can represent the same cube, a plus sign is placed in a blank square to the right of the two drawn cubes. If, on the other hand, the second drawing cannot represent the cube of the first drawing, then a negative sign is placed in the adjacent square. In the short time allowed no one attempted all items.

10. *Thurstone Flags*.—On this test two flag pictures, of the same size and of identical design, are presented occasionally in the same position, but generally in different positions. If the two drawings represent the same face of the flag, a plus sign is placed in a square on the test sheet just to the right of the two flags. If the two drawings represent opposite faces of the same flag, a minus sign is placed in the adjacent square. As in the test of *Cubes* the items were homogeneous with respect to difficulty. However, they were easy for most subjects. A few of the subjects attempted all items during the short period of time allowed.

11. *Guilford-Zimmerman Spatial Visualization*.—This is a test in which the examinee attempts to imagine the movement of a clock in space from an initial position to a final position as directed by a verbal statement. The test is divided into three parts. In the first part, one movement of the clock is required to effect the final position; in the second part, two movements are called for, and, in the third part, three movements are indicated by the directions accompanying each item. Three types of movements are required. Each type of movement refers to the revolution of the clock about an axis in one of three dimensions. The actual movement involves a revolution of the clock to the right or to the left a specified number of degrees. The word "turn" is used to designate a revolution about the base or the "6-12" axis where the numbers refer to the numerals representing hours on the clock. When the clock is tilted such that top moves either forward or backward, or in other words, when the clock is revolved about the "3-9" axis, the word "tilt" is employed. When the clock revolves about an axis perpendicular to its face, the word "rotate" is used. In the second part, two different types of movement are required, and six permutations of sequence of movement are used. In the third part, the same sequence of movements is followed in all items (rotate, tilt, and turn). Nearly all of the subjects failed to complete the entire test, but about 80 per cent attempted all items in the first two parts. Items were scored up to the point at which 67 per cent of the group attempted them.

12. *Thurstone Punched Holes*.—Each item in this test consists of a series of figures representing a square sheet of paper that has been folded by steps (as indicated by dotted lines) into smaller squares, rectangular, or triangular sizes. One or more holes are punched into the final folded form. The task for the subject is to imagine where the holes will be when the sheet is unfolded. As an aid to the subject's performance in the more difficult items one or more figures representing the appearance of the sheet of paper at intermediate stages of unfolding are presented. On the unfolded (square) sheet the subject indicates by drawing small circles where the holes will be. In the scoring of the item all holes must be properly spaced in relation to one another if credit is to be given. An item was scored right or

wrong (no partial credits were given). Nearly every subject completed all the items.

13. *Thurstone Pattern Analogies*—Adapted from similar tests in the American Council on Education series, this test is composed of items each of which consists of eight figures. The first three (stimulus) figures are labelled A, B, C, and the next five (response) figures are designated 1, 2, 3, 4, and 5. After the examinee determines the rule by which figure A is changed to figure B, he applies the rule to figure C and picks out among the five arabic numbered responses that one which satisfies the requirements of the problem. In the more complex items the examinee may frequently change his hypothesis as to the principle connecting A and B in view of limitations imposed by the nature of the five responses figures. In the time allowed, most subjects completed all items.

14. *Thurstone Form Board*—Almost identical with the *Minnesota Form Board Test*, except for the inclusion of printed instructions and a practice exercise, this test consists of items made up of several two-dimensional pieces (colored black) of various geometrical shapes which the examinee attempts to fit together in an appropriate arrangement within a larger geometric form (uncolored figure within an outline). The subject draws lines within the large white (uncolored) design to show how the black pieces can be placed in order to fit within the outline. Extreme accuracy in drawing was not required, but the solution had to be indicated clearly. No partial credits were given. Although the items became increasingly difficult as one approached the end of the test, very few subjects failed to attempt all the items in the time allowed.

The Sample

To a group of 500 male students enrolled in a two-semester course in beginning psychology at Rutgers University the battery of fourteen pencil-and-paper tests was administered. Since four class periods, spread over the last part of the first semester and the first part of the second semester of the academic year, were required for completion of the project, many of the subjects were not present at all class sessions. Makeups were given in several instances. Complete results were obtained for 360

subjects. These individuals appeared to be a representative sample of the University student body in light of biographical information obtained from each student. Consisting of 220 freshmen and sophomores and 140 juniors and seniors majoring in virtually every department of the University, the sample was deemed satisfactory. Approximately 54 per cent of the subjects were veterans of World War II. The ages of the subjects ranged from 16 to 34, the median age being 22 years.

In order that a satisfactory degree of interest might be sustained throughout the duration of the study, all students were told that they would be given their scores upon completion of testing in profile form. In fact, most subjects received scores on those tests completed during the first two class periods at the beginning of the third period. It was thought that additional motivation might be provided if the temporal interval between taking the tests and receiving the scores was not too long.

The Factor Analysis

The matrix of test intercorrelations (all product-moment) presented in Table 2 was factor-analyzed by Thurstone's centroid method in the usual manner with one minor exception (13). In the reflection of signs the criterion was that used by the workers in several of the psychological research units of the United States Army Air Forces during World War II. The algebraic sum of a column, with the diagonal entry disregarded, was employed instead of the mere number of negative signs appearing in a column. This procedure not only tends to guarantee positive sums but also appears to approximate more closely the maximizing of table totals than does the criterion involving number of negative signs.

Because of marked discrepancies between obtained communalities in the first set of centroid extractions and the estimated communalities in the diagonals, a second set of extractions was required. Following the second extraction (of seven centroid factors) the obtained communality of no test differed more than $|.07|$ from the second estimated communality.

The criterion employed for cessation of extraction of the centroid factors was also that used by workers in the psychological research units of the AAF; namely, that factoring should not

TABLE 2
Product-moment Correlation Coefficients among Fourteen Test Variables*

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
1. Verbal Comprehension ..	—	.254	.035	-.045	.075	.453	.182	-.047	.060	-.015	.145	-.002	.028	.067
2. General Reasoning254	—	.183	.132	.285	.241	.370	.135	.233	.220	.372	.308	.132	.271
3. Numerical Operations . .	.035	.183	—	.201	-.029	-.038	.261	.181	.088	.120	-.072	-.010	-.075	-.029
4. Perceptual Speed	-.045	.132	.201	—	.254	.097	.152	.517	.285	.216	.305	.277	.148	.354
5. Spatial Orientation075	.285	-.029	.254	—	.317	.247	.267	.405	.367	.609	.467	.362	.433
6. [Verbal] Completion453	.241	-.038	.097	.317	—	.227	.188	.266	.193	.402	.244	.227	.342
7. Number Series182	.370	.261	.152	.247	.227	—	.216	.284	.239	.289	.310	.303	.251
8. Identical Forms	-.047	.135	.181	.517	.267	.188	.216	—	.293	.232	.332	.265	.237	.372
9. Cubes060	.233	.088	.285	.495	.266	.284	.293	—	.355	.396	.349	.227	.367
10. Flags	-.015	.220	.120	.216	.367	.193	.239	.232	.355	—	.347	.375	.313	.391
11. Spatial Visualization . .	.145	.372	-.072	.305	.609	.402	.289	.332	.396	.347	—	.500	.370	.500
12. Punched Holes	-.002	.308	-.010	.277	.467	.244	.310	.265	.349	.375	.500	—	.328	.541
13. Pattern Analogies028	.152	-.075	.148	.362	.227	.303	.237	.227	.313	.370	.328	—	.302
14. Form Board067	.271	-.029	.354	.433	.342	.251	.372	.367	.391	.500	.541	.302	—

Decimal points omitted.

cease until the product of the two highest factor loadings is at least less than the standard error of the corresponding correlation. Such a criterion tends to yield a greater number of factors than do most other criteria. The rationale underlying this less stringent criterion is that the maximum contribution which the factor makes to the scalar product of two test vectors, or to the correlation between two tests, is no greater than the chance relationship expressed by the standard error of the correlation coefficient.

Following the completion of a set of trial rotations, it was considered advisable to extract two more centroid factors as an aid to further rotations. It was known that probably only six factors would be meaningfully identified. However, previous experience has indicated that use of additional centroid axes in the rotation process frequently brings about, more readily, a psychologically meaningful solution. The superfluous factors eventually appear as mere residuals (factors containing insignificant amounts of communality) to which no interpretation can be dependably given. Moreover, the presence of residual factors seldom interferes at the conclusion of the rotation procedure with the interpretation of those principal factors which account for most of the common-factor variance.

Fifty-six rotations of pairs of axes were required to satisfy Thurstone's criteria of positive manifold and simple structure. Each rotation was achieved graphically according to the method devised by Zimmerman (19). In general the structure determined the direction and magnitude of each new rotation. Information concerning the content of tests was put to use only toward the end of the rotation procedure when minor adjustments were made. In view of the large number of rotations the differences between the communalities of centroid factors and final rotated factors were negligible, the largest two discrepancies being .017 and .013. An orthogonal reference frame appeared to suffice for the interpretation of the factors. The final rotated factor loadings are shown in Table 4.

Interpretation of Factors

Inspection of the final rotated factor loadings in Table 4 reveals that on the whole the criteria of positive manifold and

TABLE 3
*Centroid Factor Loadings and Communalities**

Test	I	II	III	IV	V	VI	VII	VIII	IX	X
1. Verbal Comprehension	.234	-.540	.293	-.238	-.052	-.033	.103	.076	-.117	.523
2. General Reasoning	. . .	-.300	.057	.196	.168	-.133	-.085	-.038	-.063	.442
3. Numerical Operations	. . .	-.258	-.515	.100	.128	-.127	.134	.153	-.198	.485
4. Perceptual Speed477	-.363	-.259	.075	-.205	-.091	.069	.081	.546
5. Spatial Orientation213	.209	.150	-.133	-.127	.121	-.076	.054	.589
6. [Verbal] Completion	. . .	-.267	.316	-.304	-.174	.111	.134	-.031	-.063	.598
7. Number Series	. . .	-.276	-.121	.259	-.093	.089	-.182	.122	-.117	.518
8. Identical Forms204	-.340	-.275	-.058	-.066	-.134	.069	.063	.537
9. Cubes563	-.071	.029	-.052	.061	.152	-.099	.027	.372
10. Flags527	-.040	.175	.081	.122	.132	.099	.081	.407
11. Spatial Visualization726	.306	.033	-.070	-.233	-.043	-.115	-.018	.729
12. Punched Holes630	.141	.129	.203	.161	-.115	-.084	-.072	.578
13. Pattern Analogies461	.094	.108	-.232	.172	-.128	.166	.072	.392
14. Form Board662	.108	-.142	.203	.166	-.043	-.107	-.090	.611

* Decima points omitted

TABLE 4
Final Rotated Factor Loadings and Communalities*

Test	I (V)	II (V)	III (N)	IV (R)	V (V)	VI (V)	VII (S)	VIII (P)	IX (Res.)†
1. Verbal Comprehension.....	-.01	.70	.09	.17	-.01	.03	.00	-.12	-.06
2. General Reasoning.....	.24	.19	.21	.54	.00	.02	.10	.02	.01
3. Numerical Operations.....	.17	-.03	.64	.17	-.04	.01	-.04	.14	.00
4. Perceptual Speed.....	.24	-.03	.22	.01	-.01	.00	.07	.66	-.01
5. Spatial Orientation.....	.42	.07	-.04	.20	.00	-.02	.58	.16	-.04
6. [Verbal] Completion.....	.20	.66	-.02	.07	.07	.19	.28	.00	.05
7. Number Series.....	.16	.12	.34	.42	.39	.06	.17	.03	.05
8. Identical Forms.....	.24	.03	.20	-.03	.15	.06	.11	.64	.06
9. Cubes.....	.20	.06	.13	.14	.01	.18	.43	.24	.10
10. Flags.....	.15	-.08	.11	.21	.04	.26	.44	.20	.13
11. Spatial Visualization.....	.62	.20	.06	.25	.02	.06	.44	.20	.03
12. Punched Holes.....	.52	-.07	-.01	.28	.08	.36	.25	.14	-.01
13. Pattern Analogies.....	.24	.02	-.06	.09	.41	.08	.34	.15	-.08
14. Form Board.....	.53	.10	.01	.13	.00	.43	.21	.28	.02

* Decimal points omitted.

† Communalities based on rotated factor loadings expressed to three decimal places.

simple structure have been fulfilled. Six rotated factors were meaningfully identified as visualization (Vz), verballity (V), numerical facility (N), general reasoning (R), spatial-relations (S), and perceptual speed (P). Two other factors (V₁ and V₂) appeared that could not be satisfactorily defined, although their weights in certain tests were suggestive of possible interpretations. A ninth factor turned out as a residual with loadings ranging from -0.8 to +1.3.

Inasmuch as the primary purpose of the study centered about the investigation of the factors of spatial relations and visualization, the discussion relating to the identification and meaning of the other four factors will be kept to a minimum. The factors, V, N, and P are actually doublets. However, since the factorial content of the pairs of tests weighted in these three factors was well known in advance of their inclusion within the battery, there is little reason to doubt the correctness of the identification given.

It should be pointed out that the major loadings in some tests describing these three factors tended to be somewhat smaller than those reported in other studies or in manuals. This is due to the fact that many of the tests were shortened in order that they might be given within the time period available for testing.³ However, in view of the size of the sample (N = 360), loadings of .35 or greater are probably indicative of the presence of a significant amount of variance in a factor.

Somewhat greater attention should probably be given to the interpretation of the factor R. Two tests, *General Reasoning* and

³ It is possible, however, to estimate what the loadings of these three factors, as well as the loadings of the other factors, would be if the tests were not shortened (ρ). When a test is homogeneously changed in length the new factor loadings may be estimated by the formula

$$k_{mn} = k_{m1} \sqrt{\frac{n}{1 + (n-1)r_{11}}}$$

where n = number of times the test has been lengthened, or the ratio of the length of the new form to the original form;

k_{m1} = loading of factor m in the original, or unlengthened test 1 ;

k_{mn} = loading of factor m in the lengthened, or new, form of the test;

r_{11} = reliability of the unlengthened test.

If the shortened experimental forms of tests (1), (2), (3), (4), (5), and (11) are considered to be extended to their original length, the corrected loadings in the principal factor in each test are estimated to be respectively, .712, .564, .657, .673, .587, and .631, compared to the obtained loadings of .698, .537, .642, .664, .578, and .619 (which are rounded to two figures in Table 4). The assumption is made in the speed tests that the number of items completed per unit time remains constant.

Number Series, are loaded in this factor to the extent of .54 and .42, respectively. In view of the small number of items contained in the shortened form of the first test (fourteen in all) and of the consequent limitation imposed on the reliability of the test, the magnitude of first loading is substantial. Although the factor may be tentatively described as relating to some type of reasoning function, it is not clearly defined. That it may represent an ability to grasp the essential steps involved in the solution of problems presented in quantitative or symbolic terms appears to be a plausible interpretation.

Interesting to note is the fact that factor V_1 is loaded .39 and .42 in the two tests *Number Series* and *Pattern Analogies*, respectively. A highly speculative interpretation would suggest that this factor may be that of induction previously identified by Thurstone (16). When the possible existence of an induction factor is taken into account along with the fact that the test of *Pattern Analogies* received an insignificant loading of .09 in the factor R, it appears even more plausible that the factor R may represent an ability to diagnose a problem expressed in quantitative terms. If the interpretation of the R factor is correct, a significant finding is that a test (*General Reasoning*) can be constructed to measure quantitative thinking without the introduction of substantial amounts of variance in the numerical factor.

Examination of the loadings for the final rotated factors I and VII in Table 4 reveals positive, though not conclusive, evidence for the existence of two reference variables which may be meaningfully identified as spatial-relations and visualization. In short, the two hypotheses as set forth are, in the main, upheld—at least to the extent that the factorial composition of the two groups of selected tests differs.

In the following list of four tests, the first three of which were selected to test the hypothesis relating to the psychological processes involved in visualization, loadings of .35 or higher in all rotated factors including I (V_z), VII (S), and VI (V_2) may be summarized as follows:

<i>Tests</i>	Factor I (V_z)	Other Factors
(11) Spatial Visualization	.62	.44S
(12) Punched Holes	.52	.36 V_2 (.25S)
(14) Form Board	.52	.43 V_2 (.22S)
(5) Spatial Orientation	.42	.58S

In view of the presence of weights of .52 or higher in three tests *Spatial Visualization*, *Punched Holes*, and *Form Board*, (all three of which made up the group intended to represent a measure of visualization), factor I can be identified as *visualization*, even though the test of *Spatial Visualization* is loaded to the extent of .44 in factor VII (S). That the spatial-relations and visualization abilities may be required in one or more tests in either of the two groups of tests inserted in the battery was mentioned previously as a definite possibility. ✓

After taking the test of *Spatial Visualization*, many of the subjects reported that in addition to manipulating mentally the stimulus figure (an alarm clock) into the final position called for by the verbal directions, they also related the location of various parts of the stimulus object (hands, numerals, top, base, winding and setting mechanisms of the clock) to the location of corresponding parts of one or more response figures (five alarm clocks in different positions). In the easier items which required only one manipulation the role of spatial cues is undoubtedly important. On the other hand, in those items requiring two or three movements of the clock, it would appear that a greater dependence was placed upon manipulations of the clock; in fact, in the most difficult items variance associated with reasoning, verbal, and memory factors would possibly be important. However, only four items requiring a sequence of three movements were scored. Nevertheless, a small, though perhaps insignificant, loading of .25 appeared in the R factor. In short, the influence of the range of difficulty of items upon the factorial content of a test may be substantial, as a previous study has shown (6).

In two other tests, *Punched Holes* and *Form Board*, which were weighted heavily in the visualization factor, small loadings of .25 and .22, respectively, appear in the factor to be identified as spatial-relations. More important, however, are the corresponding loadings of .36 and .43 in a factor V_2 . Although not amenable to a dependable identification, this factor may be associated with the drawing (filling in) response required of the examinees. Despite their relatively high saturations in the visualization factor, these two tests appear to involve additional unknown factors.

The visualization factor loading of .42 in the test of *Spatial*

Orientation, which was chosen to represent a measure of the spatial-relations factor, is probably indicative of the use by some of the examinees of visualization. Introspective reports from subjects differed as to the technique used in working the items. The variance representing visualization ability may be attributed to the tendency of several examinees mentally to manipulate the boat, as if it were a small toy, up and down and/or left or right and to imagine concomitant changes in the scenery. Many of the subjects reported that they did not place themselves within the boat, but viewed the boat and scenery as if they were on a stationary platform some distance to the rear of the boat. One subject said that he pretended to be playing with a toy boat in a pond and to be sighting along the prow of the boat as a means of observing shifts in background scenery while he moved the boat with his hand to the right or left and/or up or down.

On the other hand, many, if not most, of the subjects pretending actually to be inside the boat, and using the prow as the guide, noted changes in background views with reference to corresponding motions of the boat. Although the test of *Spatial Orientation* appears to be weighted in both spatial-relations and visualization factors, it does seem to represent best a measure of spatial relations or spatial orientation and to vindicate its inclusion with other tests in the battery which were selected to bring out the spatial factor.

In the following list of five tests, the first three of which were chosen to yield evidence regarding the second hypothesis, loadings of .34 or higher were found in rotated factors VII (S), I (Vz), and V (V₁):

Tests	Factor VII (S)	Other Factors
(5) Spatial Orientation	.58	.42 (Vz)
(10) Flags	.44	.15 Vz)
(11) Cubes	.43	.20 Vz)
(12) Spatial Visualization	.44	.62 Vz
(13) Pattern Analogies	.34	.41 V ₁ (.24 Vz)

The magnitude of the weights in factor VII for the tests of *Spatial Orientation*, *Flags*, and *Cubes* indicates that identification of the factor as spatial relations is psychologically meaningful. Despite the substantial loading of the visualization factor

in the test of *Spatial Orientation*—a fact which has been rationalized previously—the first hypothesis regarding the psychological nature of spatial relations appears to have been upheld.

Of passing interest is the loading of .34 in the spatial relations factor appearing in the test of *Pattern Analogies*. In this factorially complex test, the presence of variance in the spatial-relations factor may have been due to the role of those changes in the design of complex figures, or patterns, which depended upon a rule involving the spatial order of parts. In the more difficult items of complex design it was usually helpful, if not necessary, to give specific attention to the spatial organization of the various geometric properties within each of the patterns appearing in the row.

A second source for possible variance in the spatial-relations factor was that of the format of each item. Pattern A and pattern B, which stood in a left-right order on the page, corresponded to the order of pattern C and one of the five alternative responses. Having been exposed to spatial tests administered earlier, many of the subjects may have transferred techniques previously learned in solving other items to the task required in the test of *Punched Holes*. Thus, the influence of mental set may have been one important reason for the appearance of the loading in the spatial-relations factor.

The results of the factor analysis seem to indicate that, in the main, the two hypotheses have been upheld. Two of the final rotated factors may be readily interpreted in terms of their weights in two groups of tests as representing the spatial-relations and visualization abilities that were hypothesized. However, the number of tests does not appear to be large enough to determine with confidence whether the abilities may be correlated to some degree.

Much needed, indeed, are other studies to yield further evidence regarding the tenability of these two hypotheses. Although two recent empirical investigations (1, 14) have indicated that similar primary factors are obtained when the same, or nearly the same, batteries of tests are administered to groups chosen under different selective conditions, it is urged that other homogeneous samples in which such variables as age, level of educational attainment, occupational classifica-

tion, and sex membership are systematically varied be employed to test the validity of the two hypotheses. Other hypotheses should be formulated regarding the psychological nature of the spatial domain and subjected to verification through use of specially devised tests and of other tests of known factorial composition. It is hoped that following more extensive research in the area of space and visualization relatively pure tests can be constructed⁴ to measure the abilities identified and that such tests can be used with others of demonstrated merit to improve materially the degree of accuracy with which numerous complex criteria can be predicted.

Summary

The primary purpose of the study was to test the tenability of two hypotheses regarding the psychological nature of spatial-relations and visualization factors. A secondary purpose was to seek to identify certain factors found in the AAF investigations with certain of Thurstone's primary abilities. Within a battery of fourteen tests, two groups of tests (three tests in each group) were included which appeared to reflect differences in the psychological processes associated with the spatial-relations and visualization abilities. In addition to the six tests expressly incorporated within the battery to yield evidence regarding the validity of the hypotheses, eight reference tests of fairly well-known factorial content were included to aid in the identification of variance found in the six tests and to answer questions of identity of the Thurstone and AAF factors.

Positive evidence for the hypotheses was to be considered attained if the two groups of tests defined separate factors and if none of the other eight tests was substantially weighted in factors unique to either group of tests. Moreover, none of the three tests in one group should contain large amounts of variance in common with tests of the other group except to the extent that a given test might consist of items that reflected the presence of that factor which was defined in the main by

⁴ Even if pure tests cannot be constructed for all factors identified in the spatial realm, means are available for attaining estimates of univocal factor scores through use of suppression tests (8).

tests of the other group. If a test did appear in one group that contained variance in the factor associated primarily with tests of the other group, a satisfactory rationalization of this finding would be required.

Product-moment correlations computed from sets of scores of 360 students in the introductory course in psychology at Rutgers University were factored by Thurstone's centroid method. Eight of these factors were rotated by graphical means to positions satisfying the criteria of positive manifold and simple structure.

In the orthogonal system six factors were identified as verbal comprehension, numerical facility, perceptual speed, reasoning, visualization, and spatial relations. In the main, the variances associated with factors identified as spatial relations and visualization were confined to the respective groups of tests initially placed within the battery to bring out the factors. In only one test in each group of three tests were substantial amounts of variance found in both the visualization and spatial-relations factors, although the larger portion of variance was in the factor common to the group in which that test appeared.

The presence of variance in these two factors was rationalized for each of the tests. Introspective reports of the subjects revealed that in many items the psychological processes used involved both spatial-relations and visualization abilities as described in the hypotheses. The range of difficulty level of test items in one test also appeared to be an important reason for the appearance of two factors.

In short, it may be concluded that the two hypotheses regarding the psychological nature of visualization and spatial relations were confirmed. However, other research projects need to be carried out with a variety of samples before a dependable generalization can be made regarding the nature of these two abilities. Since there is some evidence of still other spatial abilities (3), some or all of which may be correlated, it is recommended that a conscientious attempt be made to formulate in operational terms new hypotheses and that new tests, having been constructed in harmony with the hypotheses, be factor analyzed along with other tests of established factorial con-

tent. Once the area of space has been dependably and adequately mapped, attention can be directed toward building tests approximating pure measures of the identified abilities.

REFERENCES

1. Dudek, F. J. "The Dependence of Factorial Composition of Aptitude Tests Upon Population Differences Among Pilot Trainees. I. The Isolation of Factors." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, VIII (1948), 613-633.
2. Fruchter, B. "The Nature of Verbal Fluency." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, VIII (1948), 33-47.
3. Guilford, J. P. (Ed.) *Printed Classification Tests*, Report No. 5. *Army Air Forces Aviation Psychology Program Research Reports*. Washington, D. C.: U. S. Government Printing Office, 1947.
4. Guilford, J. P. "Factor Analysis in a Test-Development Program." *Psychological Review*, V (1948), 79-94.
5. Guilford, J. P. "Some Lessons from Aviation Psychology." *The American Psychologist*, III (1948), 3-11.
6. Guilford, J. P. "The Difficulty of a Test and its Factor Composition." *Psychometrika*, VI (1941), 67-77.
7. Guilford, J. P. "The Discovery of Aptitude and Achievement Variables." *Science*, CVI (1947), 279-282.
8. Guilford, J. P. and Michael, W. B. "Approaches to Univocal Factor Scores." *Psychometrika*, XIII (1948), 1-22.
9. Guilford, J. P. and Michael, W. B. "Estimates of Factor Loadings When a Test is Homogeneously Changed in Length." *Psychometrika*, (to be printed).
10. Guilford, J. P. and Zimmerman, W. S. "Some AAF Findings Concerning Aptitude Factors." *Occupations*, XXVI (1947), 154-159.
11. Guilford, J. P. and Zimmerman, W. S. *The Guilford-Zimmerman Aptitude Survey*. Beverly Hills, Calif.: Sheridan Supply Company, 1947.
12. Guilford, J. P. and Zimmerman, W. S. "The Guilford-Zimmerman Aptitude Survey." *Journal of Applied Psychology*, XXXII (1948), 24-34.
13. Kelley, T. L. *Crossroads in the Mind of Man*. Stanford University: Stanford Univ. Press, 1928.
14. Michael, W. B. "Factor Analyses of Tests and Criteria: A Comparative Study of Two AAF Cadet Pilot Populations." *Psychological Monographs: General and Applied*, 1949, No. 298.
15. Thurstone, L. L. *Multiple Factor Analysis*. Chicago: University of Chicago Press, 1947.
16. Thurstone, L. L. "Primary Mental Abilities." *Psychometric Monographs*, No. 1. Chicago: University of Chicago Press, 1938.
17. Thurstone, L. L. "The Perceptual Factor." *Psychometrika*, III (1938), 1-17.

18. Thurstone, L. L. and Thurstone, T. G. "Factorial Studies of Intelligence." *Psychometric Monographs*, No. 2. Chicago: University of Chicago Press, 1941.
19. Zimmerman, W. S. "A Simple Graphical Method for Orthogonal Rotation of Axes " *Psychometrika*, XI (1946), 51-55.
20. Zimmerman, W. S. "Isolation, Definition, and Measurement of Spatial-Visualization Abilities." Ph D. dissertation, University of Southern California, 1949.

ON THE USE OF INTERACTIONS AS "ERROR TERMS" IN THE ANALYSIS OF VARIANCE¹

ALLEN L. EDWARDS

University of Washington

I.

MANY psychological and educational experiments are concerned with two or more variables, each of which may be varied in two or more ways. When the variables are studied in all possible combinations in the same experiment, the experiment is said to be of *factorial* design.² As an example, let us take an experiment in which three variables are involved, *A*, *B*, and *C*. Suppose that *A* is varied in three ways, *B* is varied in two ways, and *C* is varied in four ways. Then we shall have (3) (2) (4) = 24 combinations of variables, each combination corresponding to a particular experimental condition. One replication of the experiment will thus require 24 observations and the 23 degrees of freedom available with one replication would be allocated in the following way:

<i>Sum of squares</i>		<i>df</i>
Main variables:	<i>A</i>	2
	<i>B</i>	1
	<i>C</i>	3
First order interactions:	<i>A</i> × <i>B</i>	2
	<i>A</i> × <i>C</i>	6
	<i>B</i> × <i>C</i>	3
Second order interactions:	<i>A</i> × <i>B</i> × <i>C</i>	6

If 240 subjects were available, then 10 could be assigned at random to each of the 24 experimental conditions. We would thus have 9 degrees of freedom within each of the experimental conditions or (9) (24) = 216 degrees of freedom for the varia-

¹ This paper is based upon a section of a manuscript which deals more extensively with problems of experimental design in psychological and educational research. I should like to acknowledge that I have incorporated into this paper the suggestions of Dr. Paul Horst, who served as a technical consultant on the manuscript.

² It is assumed that the reader is familiar with the treatment of the analysis of variance as given, for example, by Lindquist (6), McNemar (7), or Snedecor (8).

tion of subjects treated alike. The sum of squares for the 216 degrees of freedom would be the pooled sums of squares within groups which would be used to derive the mean square for testing the significance of the main experimental variables, the first-order interactions, and the second-order interaction.

In general, it may be said that whenever replication is present within the experimental design, the within-groups or mean square based upon replication is the appropriate error term against which to test the significance of all other mean squares. An exception to this rule, discussed in the next section, would be when the categories or classifications of one of the variables may be regarded as a *random* selection from the population being sampled.

Let us assume that in the experiment described that the *A* variable corresponds to three instructors, the *B* variable to two methods of instruction, and the *C* variable corresponds to four schools. Each instructor teaches both methods and in each of the four schools. We shall assume that 60 subjects have been selected at random within each school to serve in the experimental groups. The complete analysis of variance of achievement scores on a standardized test given at the end of the experiment would result in the following sums of squares with associated degrees of freedom:

<i>Sum of squares</i>	<i>df</i>
Instructors	2
Methods	1
Schools	3
Instructors \times Methods	2
Instructors \times Schools	6
Methods \times Schools	3
Instructors \times Methods \times Schools	6
Residual within groups	216
Total	239

Let us further assume that all of the mean squares, obtained by dividing the sums of squares by the corresponding degrees of freedom, are significant when tested against the residual mean square within groups. This would mean, first, with respect to the main variables: that significant differences are present among instructors; that the two methods differ significantly; and that there are significant differences among schools.

The interaction between instructors and methods, if significant, would mean that the differences among instructors are dependent upon the method used or that the difference between the methods depends upon the instructor variable. A significant interaction between instructors and schools would mean that the differences observed among the instructors are dependent upon the schools or that the differences observed among schools are dependent upon the instructors. A significant methods and schools interaction would mean that the difference observed between the methods is dependent upon the schools or that the differences among schools are dependent upon the method of instruction.

If the second-order interaction is significant, this would mean that the differences observed among instructors are dependent upon the methods and the schools; that the differences observed among the schools are dependent upon the instructors and the methods; or that the difference observed between the methods is dependent upon the schools and the instructors.

Now, in view of a significant second-order interaction, our conclusions concerning the main variables consisting of schools, methods, and instructors, are somewhat limited. We know that there are significant differences present for these three variables, but we know also, from the significance of the interaction, that the difference observed, let us say, for methods, is to some extent dependent upon the schools and instructors.

If our interest is only in the *two* particular methods, the *three* particular instructors, and the *four* particular schools, involved in the experiment, then our analysis and the tests of significance of the various mean squares, using the residual mean square as an error term, are appropriate. Each mean square has been evaluated and the conclusions reached are definite. Examination of the means for the various combinations of experimental conditions would probably reveal that in a particular school, one method is more effective than another, when used by a particular instructor, and we could make recommendations accordingly.

II.

In an experiment such as that described, however, our primary interest may be in the difference observed between the

two methods of instruction which we have used. Furthermore, we may wish to make recommendations beyond the particular schools investigated. Can we say that a particular method will probably be more effective, on the average, for all schools, including those we have not actually investigated?

Let us suppose that we have selected the instructors to represent particular types or personalities or abilities. The three used in the experiment are definitely not a random sample from any defined population. Nor have we selected at random from any population of methods of instruction; instead, we have picked two particular methods for investigation. But it is possible that we might have made schools a random variable by selecting the schools at random from a defined population of schools for a given city, county, or school district. If this had been our intention, of course, we would undoubtedly have taken a larger sample than the four schools at hand. Let us suppose, however, that the schools have been selected at random.

We now have the case mentioned earlier, where one of our variables may be considered a random sample from a defined population. In this sense *the schools consist merely of replications* of the experimental design in which the main variables are the instructors (varied according to type) and methods. Under this condition the highest-order interaction involving the random variable may be regarded as the appropriate error term for testing the significance of the next lower-order interactions. But before proceeding on this basis, another condition must hold true; the interaction must be significantly larger than the residual mean square within groups. It cannot, of course, be smaller except by chance. If it is smaller, the residual mean square within groups should be used in testing the significance of the next level of interactions.

Let us assume, in the present instance, that the second-order interaction is significant when tested against the mean square within groups. We now proceed to test the next level of interactions against the second-order interaction. Whichever ones of these prove not to be significant when tested against the second-order interaction may be combined with the second-order interaction to give us an error term based upon a larger number of degrees of freedom.

Under the assumptions we have made, it is quite likely that

if the second-order interaction is significant when tested against the residual mean square within groups, that some of the simple interactions will not prove to be significant when tested against the second-order interaction. The obvious reason for this is that the mean square for the second-order interaction will be larger than the residual mean square within groups. The F 's thus obtained, besides being based upon a smaller number of degrees of freedom, will be smaller than in the first instance.

Let us suppose that only the simple interaction involving instructors and methods is significant when tested against the second-order interaction. The non-significance of the interaction between methods and schools and the interaction between instructors and schools, of course, means that we no longer have any basis for inferring that the difference observed between methods is dependent upon the schools, or that the differences observed among the schools are dependent upon the methods. Similarly, the evidence would now indicate that the differences among instructors are not dependent upon the schools, or that the differences among schools are not dependent upon the instructors. The sums of squares for these two interactions may be pooled with the sum of squares for the second-order interaction, along with their associated degrees of freedom. The analysis would now take this form:

<i>Sum of squares</i>	<i>df</i>
Instructors.....	2
Methods.....	1
Schools.....	3
Instructors \times Methods.....	2
Pooled interactions.....	15
Residual within groups.....	216
Total.....	239

Now, how shall we test the significance of the mean squares for instructors, methods, and schools? If we could assume that either instructors or methods constituted a random sample from a population of instructors or a population of methods, the instructor and methods interaction might be considered an appropriate error term for testing the significance of the mean square for instructors and the mean square for methods. This, however, is not a plausible assumption. The appropriate error

term is the pooled interaction mean square based upon 15 degrees of freedom. It does include all of the interactions involving the variable which we have assumed to be randomly selected, schools. If we now test the mean squares for instructors, methods, and schools, against the pooled interaction mean square, and, if they are significant, what conclusions can be drawn?

It is the methods mean square that is of primary interest and its significance would indicate that the difference between methods was not dependent upon, or could not be accounted for, in terms of differences in the schools. A similar statement could be made concerning the instructors if this mean square was significant. In view of a significant interaction between methods and instructors, however, it would still be necessary to qualify our recommendations; the difference between the methods is still dependent upon the instructors. But the means for the various instructors teaching the various methods could be examined for whatever insight this might give us as to the nature of the interaction³

The analysis we have described is dependent upon a number of considerations and these should perhaps be emphasized once more. If the interaction or pooled interaction mean square is to be used as an error term instead of the residual mean square within groups, it should be larger than the residual mean square. If it is smaller, it is so only by chance. Furthermore, it is necessary that the categories of one of the variables in the experimental design be a random selection from the population being sampled⁴. In the experiment discussed, for example, it would be necessary for the schools to be selected at random from a defined population of schools. In this case, the categories of the randomly selected variable may be regarded as replications of the experiment, and there is some justification for the use of the

³ What if all of the first-order interactions had proved to be significant when tested against the second-order interaction? In this case, the interaction between methods and schools might be used to test the significance of the methods mean square, and the interaction between instructors and schools might be used to test the significance of the mean square for instructors. We should keep in mind that in following this procedure, our interest is in being able to generalize concerning the methods, for example, in the population of schools.

⁴ This condition will not be met by argument after the experiment has been carried through to completion. For example, it would be illogical to argue that the two particular methods of instruction selected for investigation have been randomly selected from a population of methods.

interaction as an error term, instead of the residual mean square⁶.

III.

In some complex experiments, involving many possible combinations of experimental variables and consequently many experimental conditions, replication is not used and the sums of squares for the higher-order interactions are pooled, along with the degrees of freedom associated with them, to obtain an estimate of experimental error (residual mean square within groups). The mean square thus arrived at is used in the manner in which the mean square based upon the variation within groups has been used in the experiment described, i.e., as an estimate of the uncontrolled variation against which to test the significance of the other mean squares.

An example of this design is to be found in an experiment by Crutchfield (3), in which five variables were each varied in three ways in an investigation of "behavior potentials." Animals were placed in a pulling compartment in which there was a string arranged by pulleys to a food pan. By pulling on the string the animals could pull the food pan next to the compartment and thus eat. A friction device was used to increase or decrease the force required for pulling the food pan, and behavior was studied under all possible combinations of the experimental variables.

Variable *A* was the length of the string attached to the food pan and this was varied by the use of 60 cm., 120 cm., and 240 cm. lengths. Variable *B* was the force required to pull the food pan in on the training trials and this was varied by using a low, medium, and high setting of the friction device. Variable *C* was the number of training trials given the animals and this was varied by giving 30, 60, and 90 trials. Variable *D* consisted of the number of hours between the crucial test trial and the last feeding period. This was varied with intervals of 12 hours, 24 hours, and 48 hours. The final variable, *E*, was the force re-

⁶ This is the situation in experiments involving repeated measurements on the same subjects, where the interactions involving subjects are used to provide an estimate of experimental error under the assumption that the subjects have been randomly selected from a defined population. Some of these experimental designs are described by Grant (4), Brozek and Alexander (2) and Kogan (5).

quired to pull the food pan during the crucial test trial and this was varied in the same ways as during the training trials.

By varying each of the five variables in three ways, a total of $3^5 = 243$ combinations of the variables are possible. One replication of the experiment, assigning one animal to each experimental condition, would thus require a total of 243 animals. Each additional replication would require another 243 animals. Crutchfield decided to forego any additional replications and to use as an error term a mean square based upon the higher-order interactions.

Each of the experimental variables will be based upon 2 degrees of freedom, accounting for a total of 10 degrees of freedom. The first-order interactions will each be based upon 4 degrees of freedom, accounting for a total of 40 degrees of freedom. The second-order interactions, each based upon 8 degrees of freedom, will account for 80 degrees of freedom; the third-order interactions, each based upon 16 degrees of freedom, will account for 80 degrees of freedom, and the remaining 32 degrees of freedom will be associated with the fourth-order interaction. Crutchfield pooled the sums of squares for all interactions beyond the first-order along with their degrees of freedom to obtain as his estimate of experimental error a pooled interaction mean square based upon 192 degrees of freedom.

IV.

Assumptions are involved, of course, in the pooling of the sums of squares for higher-order^j interactions and their associated degrees of freedom. In the first place, it is assumed that each of the mean squares corresponding to the higher-order interactions is an estimate of the same common population variance, i.e., the *assumption of homogeneity of variance* is involved. It is also assumed that this common variance would not differ significantly from the variance estimate obtained with replication. If the higher-order interactions are not significant—and without replication and a corresponding test of significance this must remain an assumption—then the mean square derived from these interactions will estimate the same variance as estimated by the mean square within groups.

Under these conditions, the experimental variables, *A*, *B*, *C*,

D , and E , may be tested for significance by the mean square based upon the higher order interactions. The significance of the first-order interactions may be tested in the same manner. If none of the first-order interactions is significant, this provides good evidence that none of the higher-order interactions will be significant and therefore justifies the use of the higher-order interactions as an error term.

Let us suppose, however, that one of the first-order interactions, let us say, the interaction between variable A and variable B , turns out to be highly significant. If that is the case, then the mean square based upon the pooled sum of squares for all higher-order interactions is likely to be biased in the direction of overestimating the "pure" experimental error that would have been obtained from replication of the experiment.

If the first-order interaction between A and B is significant, we should then isolate the sums of squares for the second-order interactions which involved these two variables. These second-order interactions would be $A \times B \times C$, $A \times B \times D$, and $A \times B \times E$. These sums of squares and their associated degrees of freedom would be subtracted from the pooled sum of squares and degrees of freedom for *all* higher-order interactions. Since each of the second-order interactions is based upon 8 degrees of freedom, then the subtraction of the three second-order interactions mentioned would leave a pooled sum of squares based upon 168 degrees of freedom. The significance of the three second-order interactions in question could then be tested against the residual mean square based upon 168 degrees of freedom.

It has been mentioned that homogeneity of variance of the higher-order interaction mean squares is also involved in pooling them to obtain a single estimate of experimental error. Each of the mean squares based upon a higher-order interaction might be found and the set tested for homogeneity of variance by means of Bartlett's test (1). If the test of this hypothesis does not result in the rejection of the hypothesis of a common variance, then the pooling of the various sums of squares and degrees of freedom is proper.

Although the procedure of using interactions as estimates of experimental error has been followed in much published re-

search, we should keep in mind that there is no substitute for replication. If there is an a priori reason for expecting interactions to be significant, a test, based upon replication, should be provided in the design of the experiment. If the interaction mean squares are significant, then their use as an estimate of the mean square that would have been obtained with replication, the within-groups mean square, may result in an under-evaluation of the significance of the main experimental variables.

REFERENCES

1. Bartlett, M. S. "Some Examples of Statistical Methods of Research in Agriculture and Applied Biology" *Journal of the Royal Statistical Society Supplement*, IV (1937), 137-170.
2. Brozek, J. and Alexander, H. "A Note on the Components of Variation in a Two-Way Table." *American Journal of Psychology*, LX (1947), 629-636.
3. Crutchfield, R. S. "Efficient Factorial Design." *Journal of Psychology*, V (1938), 339-346.
4. Grant, D. A. "The Latin Square Principle in the Design and Analysis of Psychological Experiments." *Psychological Bulletin*, XLV (1948), 427-442.
5. Kogan, L. S. "Analysis of Variance—Repeated Measurements." *Psychological Bulletin*, XLV (1948), 131-143.
6. Lindquist, E. F. *Statistical Analysis in Educational Research*. Boston: Houghton-Mifflin, 1940.
7. McNemar, Q. *Psychological Statistics*. New York: Wiley, 1949.
8. Snedecor, G. W. *Statistical Methods*. (4th ed.) Ames, Iowa: State College Press, 1946.

THE OBJECTIVE MEASUREMENT OF DYNAMIC TRAITS

R. B. CATTELL, A. B. HEIST, P. A. HEIST and R. G. STEWART

The Ergic Theory of Attitude Measurement

It is disconcerting that psychologists have not yet found any more objective way of measuring an individual's attitudes and interests than by asking him how strong they are. In 1935 the present writer demonstrated some degree of validity in measures of spontaneous attention and of memory, for matters of interest (3). But, apart from the work of Super (17) and one or two sporadic, incidental uses of these newer methods, the bulk of research has continued to concentrate on refinements of verbal, self-declaratory attitude and interest scales (12, 14), which, in the writer's opinion, can never satisfy the need for scientific, behavioral objectivity and meaning. Even the applied psychologists working with polls and socio-economic attitudes have regretfully had to realize that what a man says is unpredictably different from what he does and sometimes, indeed, from what he said an hour before (14). The present research, and two studies reported elsewhere (8, 9), are attempts to follow up on a more adequate scale, and to expand in new directions the original statement (3) of design for objective interest measurement.

Dynamic traits are divisible into ergs, or basic innate drives, on the one hand, and metanergs, or attitudes and sentiments, on the other (4, 5). The present study is concerned with attitudes, but, since the attitude is, in respect to modes of measurement, a prototype of all dynamic traits, the methods developed here have reference, and are applicable to, dynamic traits generally.

An attitude needs to be defined initially by five aspects, which are summarized in the paradigm:

"(1) In these circumstances (2) I (3) want so much (4) to do this (5) with that."

Here (1) defines the stimulus situation with reference to which the attitude is evoked, (2) the organism bearing the attitude, (3) strength of interest in the course of action indicated, (4) the kind of action indicated and (5) the object with which the attitude is connected. Sometimes (1) and (5) are the same.

According to the ergic theory of attitude measurement (5) an attitude may be expressed, for purposes of analysis and calculation, as a vector quantity, in which the length of the vector represents the strength of desire for (interest in) the defined course of action, and its direction represents its dynamic composition. It assumes that ergic coordinates can be discovered and defined by appropriate factor analytic procedures so that by giving the direction of the attitude with respect to these coordinates we describe the extent to which various ergs, e.g., hunger, sex, self assertion, pugnacity, gain expression through the attitude in question. An attitude is thus not regarded, by the ergic theory, as adequately expressed by the existing convention of pro- and con- an object; for an attitude about an object is far richer than a single dimension can express and is better defined in terms of all those basic-drive satisfactions which the given action to the object produces. One can, of course, correctly speak of a pro-con scale with respect to *a defined course of action*, i.e., one already defined in direction, as above. But a person may utilize the same object for many different courses of action, so that for this reason, as well as because of the possibility of fuller understanding given by expressing the ergic composition of the course of action, it is psychologically meaningless to speak of being "pro" or "con" *an object*.

The above discussion of basic theory is necessary if the meaning of the present experiments is to be understood and their findings properly applied. It leads to a formula for the strength of an attitude parallel to that used in the specification equation for expressing some particular skill in terms of primary abilities, (4) as follows:

$$I_i = S_{1j}E_{1i} + S_{2j}E_{2i} + \cdots S_{mj}E_{mi} + S_jE_{ji} \quad .$$

where I is the strength of interest of the individual i in the course of action defined by the attitude j . The S 's are the factor

loadings, which in this case we shall call the dynamic *situational indices* defining the extent to which the various ergs or drives, E_1 , E_2 , etc., are involved (for the average member of the population) in determining the course of action concerned. It is our purpose to measure I , the strength of interest, by more objective methods. The measurement of the S 's, i.e., the *directions* of the attitude vectors, is described elsewhere (5, 8). The measurement of the *strength* of an attitude is thus a measurement of *interest*. An *attitude* is measured when we measure both *interest* and *ergic composition*, i.e., length and direction of the vector.

Possible Approaches to Objective Measurement of Dynamic Traits

Considering an attitude as a dynamic trait, it is easy to perceive, from what is already known about psychodynamics, that there is a wide array of possible principles for the objective measurement of attitude strengths. The following will be briefly discussed here and the majority, those starred, will have their application to experiments described precisely.

A. Criterion Methods, (a) Interactive.—By these are meant methods of measurement too long and difficult for routine test use, but which, when properly applied (19), supply data that can be taken as a true measurement of what is meant by interest—in objective, "interactive" (4) units—in the real life situation.

* (1) *Money*. Fraction or absolute amount of the individual's income that he spends on certain courses of action.

* (2) *Time*. Fraction of the individual's time that he gives to certain courses of action. (18)

B. Criterion Methods, (b) Solipsistic.—By these are meant methods of measurement dependent on introspection and self assessment but which, in the specially controlled circumstances of experiment with intelligent, cooperative subjects, can be used as criterion data.

* (3) The classical "opinionaire" method, as used by Thurstone (20) and others.

* (4) The "preference" method, in which the individual is presented with alternate courses of action (attitudes) and asked which he would prefer to satisfy. This is

done in all possible paired comparisons among, in this case, 50 attitudes, and thus supplies a more thorough, pointed measure along the lines of (3). It is the same situation for human beings as that presented to animals in the classical "choice box" experiment on motivation strength (21) except that the reaction is a verbal only.

C. Attention-Memory (Learning) Methods; (a) in the immediate situation.—These depend on the principle that interest (incentive) is a determiner of attention, rate of learning, inhibitory effects on other processes, etc., and seeks to measure interest through such effects.

- (5) *Attention time* Recording the length of time or the rank order in which the individual will spontaneously attend to various stimuli.
- * (6) *Immediate Memory.* Since there seems little point, as far as we know, in separating measures of "observation" from "immediate memory," this records instead of "attention" the amount of various interest data *recalled* almost immediately after exposure. As indicated later, the measure was tried separately for statements facilitating the expression of the attitude and statements frustrating it.
- (7) *Reminiscence.* It would seem likely that reminiscence, the selective action of memory as determined by contrasting immediate with more remote recollection, might be particularly correlated with interest.
- * (8) *Distraction.* This method aims at measuring the attention effect indirectly by recording the *failure* to perceive surrounding material when the interesting object is presented.
- (9) *Retro-active Inhibition.* As with distraction, the interest an individual has for certain matters, particularly in the deeper interests, might be validly measured by the amount of retro-active inhibition their consideration exerts upon some prior, standard learning process.

D. Methods Appraising Cognitive and Dynamic Structure due to Interests.—The methods under C depend on learning effects of interest *in the immediate test situation*, but if we are willing

to accept the slight error due to time lag we can measure interests alternatively by the effects they have had *in the course of time* upon information, skills and dynamic response habits.

* (10) *Information*. This method tests the individual's information about facts, devices, etc., necessary to *implement the course of action* in which he is interested (not necessarily knowledge about the object).

* (11) *Speed of Decision (Reaction time)*. This method assumes that decisions will be given more quickly for questions in regard to which the individual has more intense conviction. Preliminary work already indicates the probability of this.¹

(12) *Level of Skills*. The extent of the built-up skills in a certain course of action may, like the level of information, provide a measure of the strength of interest therein, e.g., performance on a piano provides an index of musical interests, or skill in shooting of hunting interests. Time and errors in suitably chosen diagnostic performances would thus provide a measure of this area. So also might speed of decision in a different context from (1) above, namely in that there would be, through practice, greater quickness in making decisions in those fields with which *S* is familiar.

E. Autism Methods. In research on so-called "projective" tests the present writer has pointed out (7) that devices in this area are more aptly called apperception tests (since such measures include both cognitive and dynamic sources of distortion). Within the apperceptive class, however, we may distinguish *autism tests*, which deal with *distortions of perception, reasoning and memory through dynamic traits* alone. *Ego defense dynamisms tests* are a sub-category within *autism tests*. The autism methods

¹ Chant and Salter (10), presenting an "attitude to war" opinionaire to a group of mainly pacifist subjects, found that items which demanded *longer* decision had a larger P.G.R. ($0.72 \pm .07$), but that more "militaristic" items had larger P.G.R. and more neutral items a longer decision time (i.e., curvilinear relationships exist). What bears more simply on our approach is their finding that *rejected statements* had larger deflections ($.71 \pm .16$) and longer decision times. (Mean $2.6 \pm .08$ greater than accepted).

At the reading of the present paper at the annual Mid-Western A.P.A. meeting in Chicago 1949, Gallenbeck (13) announced that he had results, but more finely analyzed, entirely confirming the relation between affirmative decision times and strength of convictions, presented here. The results of Postman (15) are also in agreement with this use of decision time as a strength of attitude measure.

used to measure the dynamic traits of special significance to personality are obviously applicable to interests in general, though the defense dynamism tests are not so relevant.

- * (13) *Misperception (Perceptual Autism or Illusion)*.—In this method defective sensory presentations (mainly of words) are made such that the individual may be tempted to apperceive them in accordance with his wishes. He is scored on the number misperceived to fit in with his attitude.
- * (14) *False Belief (Reasoning Autism or Delusion)*.—The method presents a number of manipulatable statements of fact and logic so chosen that the individual with a strong attitude will experience a need to distort his factual beliefs in a certain direction better to support his attitude.
- * (15) *Phantasy*. This method treats phantasy *in toto* and not merely the defense dynamism forms. A measure of *time* spent phantasying or of *choice* of phantasy reading in presented alternatives is recorded.
- * (16) *Projection* (Defense dynamism). Two types of controlled, selective answer tests are possible in this area. (a) That in which the picture or the verbal statement of activity is fixed and the subject selects the best of the alternative dynamic "explanation" of the behavior (See design in (9)). (b) That in which the subject chooses the activities, from a presented list, of which he prefers to "explain" the motive. The latter is psychologically more complex but has not been tried and it was the especial interest of one co-worker to try it out here. (9)
- (17) *Ego Defense Dynamisms*. It is possible that any other defense dynamism, e.g., reaction formation, identification, rationalization, true projection, defensive phantasy could be used here, by methods described elsewhere (7), but such methods would be restricted by applying only to interests connected with ego conflicts and were not tried out at this stage of exploration.

F. Activity Level Methods, (a) Psychological.—In this cate-

gory, which includes some relatively miscellaneous approaches, we include attempts to measure increases in the general excitement level of the organism due to arousal of interest by the stimulus in the experiment.

- * (18) *Fluency*. A measure of the sheer amount written, in a given time, in a "completion" test of statements concerning a given attitude.
- * (19) *Speed of Reading*. A method based on the hypothesis that an individual will read more rapidly material which interests him and which is in agreement with his own attitudes.
- * (20) *Work-Endurance Measures*. This method plans to measure work output (endurance of fatigue) or endurance of pain or discomfort in the interest of various attitudes and is thus analogous to the obstruction method in animal motivation studies (21). Miniature situations involving satisfaction of the particular attitudes could be made, for example, in terms of satisfaction of curiosity in reading about facts contributory to the total attitude satisfaction.

G. Activity Level Methods, (b) Physiological.—The known, promising methods of measuring increase in activity level are greater in the physiological field, where autonomic and metabolic measures have been more developed.

- * (21) *Psychogalvanic Response*. The percentage decrease in resistance was measured on exposure of statements favoring and opposing the given attitude.
- * (22) *Pulse Rate*. Difference of rate before and after presentation of stimulus defining attitude.
- (23) *Metabolic Rate*. A better measure, to which the above is only an approximation, would be the increase in metabolic rate following, in a discovered optimum period, the presentation of the attitude statements. Because of technical difficulties we had to be content with (22).
- (24) *Muscle tension*. There is evidence in the work of Duffy that general muscle tension is as sensitive and reliable a measure of conation as is the P.G.R. For lack of further work confirming the measurement of conation

by this method, however, we eventually did not use *general* tension, but (25) below.

- * (25) *Writing Pressure*. The subject was asked to write "Yes" or "No" according to his reaction to presented attitude statements. A device beneath the writing desk measured the handwriting pressure he exerted in these responses.

Twenty-five distinct methods of objective attitude measurements are suggested, above, to be of promise; but nine of them—(5), (7), (9), (12), (15), (17), (20), (23) and (24)—were not tried in the present experiment, some because of special technical difficulties, some because of similarity to methods already in the sample and some, namely (5), (15), and (17), because an idea of their effectiveness has already been gained from earlier research (3), (7), (11). Of the sixteen methods tried, twelve are described here and the rest elsewhere (9).

The Experimental Design

The proof of goodness of an attitude measurement method is valuable only if it applies to any kind of attitude. Consequently, it was our objective to design the experiment so that a wide range of methods could be applied to a sufficient sample of a wide range of attitudes. Twelve attitudes were taken, sampled from (1) those of massive importance in everyday life (and therefore of interest to clinicians), from (2) those sampling distinct basic drives and (3) those of different social and intellectual interest areas (such as have been of interest to social psychologists). The list was based mainly on the fifteen categories of Cattell's *Interest Test* (6).

The twelve attitudes chosen for experiment with the various measurement methods here described were actually administered to the group in a total set of fifty attitudes, in connection with an experiment described elsewhere (8). This inclusion in a large group gave certain advantages, notably, that the preference score could be the rank order in fifty attitudes rather than in twelve. The twelve attitudes are set out below according to their index numbers among the fifty (8).

- (1) I want to play more indoor sociable games, such as card games,

- (2) I want to spend somewhat more on drinking and smoking than I am now able to do.
- (6) I want to become proficient -if possible to excel my colleagues—in my chosen career.
- (10) I want more time to enjoy sleep and rest.
- (11) I want to listen to music.
- (16) I want to know more science.
- (19) I want to see organized religion maintain or increase its influence.
- (22) I want to attend football games and follow the fate of teams.
- (30) I like to see a good movie or play every week or so.
- (34) I want to get my wife the clothes she likes and to save her from the more toilsome household drudgeries.
- (36) I want to be smartly dressed, with a personal appearance that commands admiration.
- (44) I want to feel that I am in touch with God, or some principle in the universe that gives meaning and help in my struggles.

Upon these twelve attitudes the twelve methods of measurement set out below were tried. Four methods—(4), (6), (10) and (21)—were tried on *all* attitudes; two methods—(1) and (2) — were tried on seven attitudes; and the remaining, newer methods were tried on one attitude each.

*Brief List of Methods Examined Here (Entirely New
Methods in Italics)*

- (1) Money expended
- (2) Time expended
- (4) Preferences
- (6) Immediate Memory
- (8) *Distraction*
- (10) Information
- (11) *Speed of Decision*
- (13) *Misperception (Illusion)*
- (14) *False Belief (Delusion)*
- (18) *Fluency*
- (19) *Speed of Reading*
- (21) Psychogalvanic Response

It was our aim to measure validity in terms of *correlation with the pooled result of all methods*. But from existing information it is likely that some methods are better than others and, indeed, six of the above methods, those in italics, are "long shots," with no previous work on them whatever; so we decided to make the validating *core* out of the first six—hereinafter designated "tried" tests, because previous work has shown (1), (3), (11),

(16), (17), (18), (19) some degree of validity. Also, we desired to know the relative goodness of these first six *tried* tests with greater accuracy, whereas we were interested only as to whether there exists any validity *at all* in the *exploratory* (italicised) tests. It was for this reason that the *tried* test methods were applied to the majority of attitudes, but each of the exploratory methods was tried on one attitude only.

The subjects were a population homogeneous as to sex (men) and chosen to have family interests (all were married) but otherwise diverse (some students, some business men) and ranging in age from 20-40 (80 per cent between 25 and 33) so that though all possessed the attitudes in question they would do so in diverse degrees. Six methods (the "tried" methods) were applied to all subjects but not on all attitudes, for each 40 subjects took a different pair of attitudes. The six exploratory methods were therefore each applied only to one attitude and 40 subjects.

A more detailed statement of the method of administration of the twelve methods follows.

(1) *Money Expended*—(No. 1 in general list; used on all attitudes.)—Two weeks a month apart and clear of any special holiday season, were taken and *S* was asked to record his expenditure on the particular interest activity concerned for the whole week. Reliability coefficients were calculated with respect to the two-week periods.

(2) *Time Expended*—(No. 2 in general list; used on all attitudes.)—In the same two weeks *S* recorded separately for each and at the time the number of hours spent in the given activity interest (See (18).)

(3) *Preference*—(No. 4 in general list above; used on all attitudes.)—A matrix of cells was constructed, constituted by the triangular area bounded by the full fifty attitudes arranged in rows on the right and in columns from right to left. Each cell thus represented a possible comparison of the strength of one attitude with that of another. *S* thus made 1225 paired comparisons, indicating in each case which of the two attitude goals concerned in the comparison he would rather satisfy. The score for a given attitude was the fraction of the 49 comparisons in which it was the preferred member.

(4) *Immediate Memory*—(No. 6 in above list; used on all

attitudes.)—A series of 500 brief statements, 10 to an attitude, equally divided among those pro and con each of the attitudes, were presented tachistoscopically at 6-second intervals. They were presented in a series of 42 discs, each consisting of 12 statements randomly mixed from among the fifty attitudes. As examples of the five pro or facilitating and five frustrating stimuli used in connection with each attitude we may take from attitude 6 (wanting success in one's career) the two statements "Success in career assures happiness" and "The successful careerman is always selfish." *S* was told at the beginning that after every 12 statements (and a pause of 25 seconds) he would be asked to recall, in 30 seconds, all that he could remember of "the phrases, statements or ideas presented in the last period." Credit for recall was given when the essential idea of the item was re-iterated regardless of verbal form. This same situation and set of attitudes was used simultaneously to get the P. G. R. responses described below.

(5) *Distraction*—(No. 8 in above list; used on attitudes 36 through 40.)—Statements similar to the above were exposed, ten to each attitude but intermixed. *S* was told he would be given 10 seconds to look at each statement and might be asked to repeat it (he *was* asked intermittently) as well as to recall the nonsense syllables scattered around the statement. Twelve or thirteen nonsense syllables were in the margins around each statement. *S* was given 10 seconds to write down above recalled items.

(6) *Information*—(No. 10 in general list; used on all attitudes.)—Ten information items, each with multiple-choice selective answers, were presented for each attitude. The information dealt, not with the *object* (which would measure total interest in the object) but with knowledge required in following *the course of action* connected with the attitude. *S* was asked to leave no item unanswered but to guess. Scored on total number right. A typical example may be taken from attitude 22, on wanting to follow football games as a spectator:

"In the {Orange
Sugar } Bowl game of January 1st, 1948 {Georgia
Cotton } {Michigan
S. M. U. }
defeated {Alabama }
{California}."

(7) *Speed of Decision*—(No. 11 in general list; used on attitude No. 1.)—Ten questions were presented for each attitude. They were chosen to be such that all *S*'s would give *some* degree of affirmative answer, and *S*'s were told to give an answer in the form "Probably," "Yes" or "Certainly," i.e., definitely and emphatically yes. For example, "Do you want the sale of liquor to children to be prohibited?" This uni-directional response was necessary because previous research has indicated (15) that a short decision time is associated both with very affirmative and very negative responses. We need a question such that reaction time would work only in one direction.

(8) *Misperception (Illusion)*—(No. 13 in general list; used on attitude No. 2.)—Ten attitudes statements, positively expressing the attitude, were presented for each attitude. *S* was instructed to expect 1 second tachistoscopic exposures of sentences, to repeat what they said and to note any misspellings. Sentences were such as "I want to eat a chocholate sundea," "I want to reduse my weight throuogh work." Ten statements not connected with any dynamic need were presented as a control on *S*'s normal carefulness of spelling perception.

(9) *False Belief (Delusion)*—(No. 14 in general list; used on attitudes 41, 42, 43 and 44.)—Ten statements for each attitude were presented *S* as an "Information Test." The five multiple-choice alternative factual endings to each statement were such as to give greater or less factual support to the attitude *S* might desire to maintain. Thus on attitude 44, "During the war church attendance increased greatly and since V-J day it has (declined slightly; tended to increase still more; stayed at its high peak; returned to its pre-war level; fallen to its lowest point since 1920).

(10) *Fluency*—(No. 18 in general list; attitudes 31, 32, 33, 34, 35.)—*S* was shown each of the ten statements originally used to express each attitude and was told to write as much on the topic of each as possible in 1 minute. It was noted that this 'fluency' increased slightly but steadily in successive attitudes, so *S* was run through attitudes in both direction. At this administration no check was kept of relative fluency on pro and con statements. Score was total number of words produced.

(11) *Speed of Reading*—(No. 19 in general list; tried on attitudes 14, 15, 16, and 17).—Six statements were presented

for each attitude, three favoring the attitude and three against it, but in random order. *S* was timed on reading statements aloud, the negative item speed being subtracted from the positive on the assumption that *S* would read more rapidly those statements which expressed his desires.

(12) *Magnitude of Psychogalvanic Response*—(No. 21 in general list; tried on all attitudes). The P.G.R. was applied with the technical conditions described in earlier work by the senior author (2), the deflection being measured in percentage of the absolute resistance. For each attitude the deflection was taken to tachistoscopic exposures of five statements favoring the attitude and five opposing it, the instructions and exposed statements being those used in the *Immediate Memory Test*.

TABLE 1
Reliabilities of "Tried" Methods

	Attitude Number*												Mean thro' Z score
	1	2	6	10	11	15	19	22	30	34	36	44	
(10) Information	.64	.39	.13	.20	.86	.14	.68	.92	.90	.31	.36	.50	.59
(21) Psychogalvan	.12	.31	.84	.98	.96	.84	.88	.93	.96	.80	.70	.63	.85
(4) Preference	.70	.89	.88	.88	.88	.92	.96	.87	.67	.90	.92	.98	.90
(6) Immed. Memory	.32	.13	.39	.47	.53	.21	/	.74	.86	.47	.21	.44	.50
(2) Time Exp.	.92	/	.38	.96	.97	/	.98	.99	.71	/	.94	/	.94
(1) Money Exp.	/	.86	.48	.98	.96	/	.99	.99	.96	.84	.67	/	.94

* These correspond to the numbers in the complete description of fifty attitude n (8).

Scoring was carried out for facilitating and frustrating sets separately and also for all together, as discussed below.

Results

As indicated above, the measurement of each attitude was split wherever possible into two sets of five items, in order to get a reliability; but, where the measures had first to be split into pro and con items, the reliability was reduced to two items against three.

The reliabilities for test forms applied to all twelve attitudes and corrected to 10-item length are as shown in Table 1.

For Immediate Memory with unfavorable statements (Attitudes 6 and 10) the reliability was .32; for facilitating statements, .45; for the Distraction measure (Att. 36), .64; for Speed

of Reading (Att. 16), .79; for Misperception (Att. 2), .43; for False Belief (Att. 44), .53, for Fluency (Att. 34), .68; and for Speed of Decision (Att. 1), .90. Apart from the methods of comparing the speed of reading of favorable and unfavorable views and the method of misperception of spelling, therefore, any failure of a method to attain recognizable validity cannot be imputed to any large extent to unreliability of the tests. These two methods, as well as the immediate memory method, however, evidently need improvement in items and procedure, to gain reliability sufficient for a more exact appraisal of validity. Information and P.G.R. could also be improved on certain attitudes which offer specific difficulties in test item design. For example, Attitude 10, "I want more time to enjoy sleep and rest," evidently makes severe demands upon the

TABLE 2
Validities of "Tried" Methods

	1	2	3	4	5	6
1. Time Exp						
2. Money Exp47					
3. Preference21	.25				
4. Information16	.15	.16			
5. Immed. Memory01	.03	.13	.01		
6. Psychogalvan08	.04	.03	.04	.15	
Row 1 Mean Validity in regard to all19	.19	.16	.10	.05	.07
Row 2 Mean for 1st four methods28	.29	.21	.16	.02	.05
Row 3 Mean Reliabilities94	.94	.90	.59	.50	.85

experimenter's subtlety in choosing information items connected with this interest, for the ten items used attained a split half reliability of only .20.

For the six methods used on all twelve attitudes, twelve correlation matrices were worked out, and averaged (cell by cell), by Fisher's *Z* function, to give the values in Table 2 for the mean intercorrelation of the different methods applied to a representative set of attitudes.

No factor analysis has been attempted on so few variables, but what is substantially the loading of each method in the first general factor has been indicated by averaging its correlations with all other methods. This "internal validity" we shall take as the best basis for deciding the relative validities of the various methods.

The calculation of the standard error on these correlations is somewhat complex. Each r in the body of the matrix is the mean of eight to twelve r 's on 40 men each. Since they are averaged through Fisher's Z function the standard error of each would have $\sqrt{N-3}$ in the denominator, so that the standard error of the mean would be equivalent to an r on a population of between $(N-3) \times 8$ and $(N-3) \times 12$, i.e., 296 to 414. However, the validation r 's are each the mean of five r 's each with the above standard error. The fact that the five latter represent independent experiments but not independent groups creates some difficulty, but assuming independence through experiment and applying the $N-3$ denominator we arrive at from 1480 to 2070 cases as the population on which the r 's in Row 1 are based. On this basis the validities of methods 1, 2, 3 and 4 are

TABLE 3
Reliabilities and Validities of Exploratory Methods

Method	Attitude	Reliability	Method of Preference	Method of Information	Method of Money or Time & Money	Mean of all
Speed of Decision	No. 1	.90	.09	.26	.16	.16
Distraction	No. 36	.64	.29	.35	10 & .08	.18
Misperception (Illusion)	No. 2	.43	.00	.13	.01	.06
False Belief (Delusion)	No. 44	.53	.33	.11	.10	.18
Fluency	No. 34	.68	.25	.01	.03	.08
Speed of Reading	No. 16	.79	.11	.06	.00	.05

significant at the 1 per cent level, 6 at between 1 and 5 per cent level and 5 barely at the 5 per cent level, though its correlations are consistently positive.

The results for the six newer methods are set out in Table 3 which shows, first, the reliability of the measurement and the attitude (Numbered as above) upon which it was tried; second, its correlations with the best four methods (1, 2, 3 and 4) above, and last, its mean correlation with all methods tried, usually six.

Speed of decision, false belief and distraction are the only methods in which the pattern of correlations indicates some validity (at the 5 per cent level). Evidently the finding of Bruner (1) that misperception effects can arise from attitudes is one which shows up in differences of means but is not strong or constant enough to show up in the more exacting examination by correlations and with methods of this kind.

Speed of reading seems unrelated to agreement with the views read and there is only a faint suggestion that fluency is related, though both show their highest correlation with the best method, namely Preference. These and the other newer methods are being tried out again, each on *ten* attitudes, since the peculiarities of a single attitude, as in the present research, may give an unfair impression.

Certain possibilities in both the more basic and the more exploratory methods remain to be examined, notably (a) the possibility that higher validities will be found in ipsative (4) than with normative scoring, (b) the possibility that some relations are curvilinear, (c) the possibility that there are contrasting effects not only between stimuli that have to do with an attitude and those that do not, but also between those that favor and those that frustrate the attitude.

It will be remembered that ipsative scoring expresses the score relative to some average or total of the given *individual*, whereas normative scoring expresses it relative to the distribution in the *group* (4). Where the raw score expresses some real interaction of the individual with his environment—some behavior that may be considered a real function of interest, as the tests of information, time and money expenditures, etc., do—the present figures were scored normatively, i e., in standard scores, before correlating. Preferences, the P G R. and the Immediate Memory tests, however, were scored ipsatively, for in the last case, for example, the immediate score is clearly relative to the individual's standards. His intelligence and memory may be such that he exceeds the score of another person on a particular attitude even though his interest in that attitude is quite small. In the second case individual physiological differences in reactivity (one person may have an average P.G.R. deflection five times as large as another) need to be corrected. The first method, preferences, is automatically ipsative in scoring, since each person has the same total.

This is no place to attempt a discussion of the ipsative-normative scoring problem, which, however, must be recognized as peculiarly insistent in the field of interest measurement and has, for that reason, been fully discussed in a first approach to the theory of interest testing (3). There is as yet no simple

solution and indeed a claim can be made for putting almost any interest measure on an ipsative basis before putting it into normative scores. For example, the extent of the need expressed in a money expenditure can only be properly gauged when we know how much money the individual possesses. However, in this dilemma we have thought it best to turn to ipsative scoring only when the individual differences in mean scores are patently great and when there are good reasons for believing that some personal constant, e.g., physiological reactivity or general power of immediate memory, mediates strongly between the behavioral expression of interest and the particular manifestation we have chosen to test.

No digression comparable to the above will be taken into curvilinearity. Suffice it that one investigator (15) has shown that speed of decision is related to strength of conviction in bimodal fashion, a quick decision being made where attitudes are strongly for or against the question. A similar complexity has been found on the relationships of P.G.R. response and memory value (16) and P.G.R. response and speed of decision. (10)

However, in our correlation plots we have encountered no persistent curvilinearity, and with the exception of suggestions thereof in speed of decision, P.G.R. response, fluency and speed of reading, which require further investigation, we believe that there is no measurement problem in this respect.

On the other hand, the problem of differences between the effects of statements favoring the successful expression of an attitude and those frustrating it is a very real one for certain methods, and in one method, the P.G.R., we had reason to believe that the poor validations obtained were due to the neutralization of two conflicting significant responses. Our search in this direction was stimulated also by the finding of Whately Smith (16) of a curvilinear relation between memory value of words and their P.G.R. deflection, such that the largest deflections were found both with words very well remembered and words very poorly remembered.

Consequently, in the ten items exposed both for the P.G.R. measures and for the *Immediate Memory Test* five were made "facilitating" and five "frustrating" items for P.G.R. and Im-

mediate Memory were separately scored and correlated. Owing to the complexity of the inter-relations and the lack of significance of some of the results only the positive indications will be briefly set out, as follows:

Attitudes evoking larger deflections on facilitating items tend to have also larger deflections on frustration items than do other attitudes ($r = .29$ and $.36$) and the same occurs to a lesser extent in immediate memory measures.

Larger deflections on facilitating items in an attitude are associated with poorer immediate memory for that attitude, particularly in its frustrating items ($-.25$ and $-.36$). The implications of the last statement, together with the Whately Smith findings, are clearly that both immediate memory and the P.G.R. have a more complex relation to interest than the simple linear one hoped for in this exploratory study. The bearings of this on further research are discussed below.

Discussion

Some observations not reducible to the above statistical digest need first to be added. These concern mainly the operation of particular methods and can be presented seriatim.

It was the general opinion of the experimenters that the reliabilities obtained for the expenditure of time and money methods were higher than the true dependability of the observations warranted. Subjects, on close examination, were found to have been careless about their records of actual expenditures and to have made guesses, the similarity of which in the two weeks in question raised the apparent reliability. It is suggested that in further, more intensive experiments these records be kept in more detail and over longer periods than one week. In two attitudes, notably that dealing with expenditure on the wife and among students with very restricted means, some experimenters noted a curious tendency to inverse relationship between the amount spent and the stated intensity (Preference score) of interest. This general problem of the tendency of conscious, verbal intensity to be related to the extent of the *frustration* of the need rather than to the basic amount of need satisfaction occurring in the given attitude justifies special investigation.

In the *Immediate Memory Test* the impression of experimenters while administering it was that it was not working very well. The usual positional effects were noticed (first and last in each run of 12 being best remembered) but these were cancelled as far as possible by giving each attitude an equal positional chance. Since briefer items were apparently more frequently remembered it is suggested that future work attempt to bring all stimulus statements to five-, six- or seven-word length. There is some evidence, additional to that implicit in the above correlations, that good validity could be obtained for a memory method concentrating on failure to remember *frustrating* statements. In one attitude r 's of .40 with Preference and .14 with Time and Money were obtained for this "memory failure with contrary statements" score.

Both in the memory test and in the P.G.R. some distortions were produced by items which were unintentionally embarrassing or amusing, and subjects were suspected of not repeating the former even though they remembered them. Experimenters also suspected that dynamic effects, both in memory and P.G.R., tended to spread from a particular item to the items that happened to be neighbors. Some of the pooriness of validity of the P.G.R. test was believed by most experimenters to arise from purely technical difficulties, e.g., change of meaning of the size of deflection with different absolute resistances, so that improved apparatus, such as the self-recording and more accurately balanceable instrument since constructed, is expected to yield validities equivalent to the other methods. It is also suggested that one or two "buffer" items be introduced before each run of a dozen or so stimuli, since it was noted that the first items after an interval tended, regardless of significance, to produce appreciable deflections.

However, the use of the P.G.R. and Immediate Memory Methods can never be satisfactory until the problem of the relative significance of responses to "facilitating-frustrating" stimuli, involving the above mentioned Whately-Smith effect, has been cleared up. The senior author believes that the current use of the P.G.R. could best be improved by using solely noxious (a specific variety of frustrating) stimuli and counting the response as a true function of the strength of the attitude threatened.

Improvement of the promising 'Distraction' test is suggested through employing memorizing material more finely divisible and easier to remember than nonsense syllables. Numbers would be one such medium.

In the equally promising Speed of Decision method it is possible that some useful compromise method might be worked out in which the extent of the subject's stated agreement or disagreement would be taken into account as well as his decision time. This would bring the advantage that questions inviting *negative* answers could also be used and the experimenter would not need to strain his ingenuity seeking questions that admit only of various degrees of positive answer. The relation of decision time to degree of positiveness found in this method (for attitude No. 1) is shown in Table 4.

Although the above relation might not represent a correlation of more than .10 or .20, the combination of a speed score

TABLE 4
Relation of Speed to Positiveness of Decision

	Probably	Response	
		Yes	Definitely
Times response given for 40 subjects	315	442	443
Average seconds per response	2.9	2.2	1.7

with a degree-of-assent score should reach an appreciably higher validity.

So much for special methods. In the experiment as a whole the chief weaknesses resided in: (1) the great demand on the subject's time, which tended to produce fatigue and boredom inconsistent with good cooperation; (2) the multiplicity of experimenters (seven different people in various aspects of the undertaking); (3) the defectiveness of individual test items, notably in the Information, Immediate Memory and Misperception tests, due to absence of item analyses.

The first is unavoidable, except with expensively hired subjects, if many methods are to be cross-validated in a widely planned exploratory study, but need not interfere in the more restricted local studies that can now be carried out with the knowledge here presented as to the general field. The second may be a blessing in disguise: if a method is such that it yields

valid results in the hands of several experimenters one may be sure that it is a well-defined method and one valid in many circumstances. The third raises the general problem of whether item analyses should be carried out before or after the validity of a certain type of test has been established. The writers believe that in exploratory studies the items should be designed on a sufficiently clear general principle. If this proves to have *any* validity the less valid items can later be combed out by item analysis (consistency with the test as a whole).

The above considerations may indicate why the validity coefficients of some methods have been called "acceptable and promising," even though the correlations, significant at only the 5 per cent level, are still short of what would normally be considered good validity. Our first aim was an exploration to discover new methods of *any* real validity. The second aim, of improving them to *practicable* validity, can be predicted to encounter difficulties in certain cases. For when corrected for attenuation by low reliability the correlations with the criterion for most of the above methods still hover only between 0.3 and 0.5, and we accept the position of Guilford that in psychometry validities below 0.5 are not of much practical use.

However, the improvements indicated above are likely to raise the validity more than the reliability, and it is, moreover, possible that the present reliabilities, as indicated above, are overestimated for certain tests. Nevertheless, even if it be supposed that the validities of the separate methods could never be raised above 0.5, a very acceptable and effective battery could be made from a combination of half a dozen of these methods. For apparently only to a slight degree accounted for by error, what is the specific element in each? Most likely it is a combination of (a) other dynamic traits partly determining interest in the specific items chosen to represent the attitude, (b) individual abilities and temperamental qualities affecting the given medium of measurement, e.g., power of memory in the memory test, autonomic reactivity in the P.G.R., (c) life circumstances which cause certain expressions of the attitude to be unused or inhibited in certain persons.

There is obviously much scope for research here, both on the sources of chance error in our measurements, i.e., on determin-

ing the physiological, instrumental and other causes of low reliability of the dynamic measurement, which, for the moment, we have brushed aside as "chance error," as well as on the more systematic specific factors discussed in the last paragraph, but our interest at this stage has been to pursue the element of real validity, leaving the causes of non-validity for later examination—wherever some validity is found

These last considerations raise a question to which both space and the roughness of data compel us to give only a tentative answer here. By taking validity as the mean correlation with the pool, we have implicitly assumed that only a single general factor is of importance. There is, however, some indication of a less clearly developed block of intercorrelating methods, additional to the main block, including time and money expenditures, preference and information. It shows itself best in the correlations for one or two particular attitudes (6 and 10) where a significant cluster appears in Immediate Memory (failure to remember statements contrary to the attitude), Preference and Projection (averaging .36 and .17 in the respective attitudes) test and slightly in Information and P.G.R., but scarcely at all in time and money expenditures or memory for favorable, facilitating statements. This may be that special aspect of an attitude strength represented by unsatisfied drive, but until further studies confirm the pattern discussion would be premature.

Conclusions

1. From the administration of tests of attitude strength ("interest in a defined course of action") involving twelve different methods, applied to most of twelve different attitudes, the mean reliability of each method and the mean correlation of each method with the other methods was obtained.
2. The reliabilities varied from moderate to good, but only eight of the methods had validities that were significant
3. The validities were defined as the mean correlation with a pool of four or six "tried" methods, which were set aside at the beginning as psychologically sound criteria and of some tested worth. These were.—Expenditure of Money, Expenditure of Time, Stated Preference in Paired Comparisons, Information

Implementing a Course of Action, Immediate Memory and Psychogalvanic Response to statements concerning the attitude. Only the first four of these reached incontrovertible validities.

4. The comparative failure of Immediate Memory and the P.G.R., despite good previous indications, seems traceable to complex relations, notably the Whately-Smith effect, differentiating the Memory response and the P.G.R. response (but in different ways) respectively to facilitating and frustrating verbal stimuli.

5. In several methods where the interest response is mediated by the extent of the individual's possession of some secondary personality factor large differences appear between the average magnitudes of the individuals' mean responses to *all* attitude interests and it is then necessary to rescale the score ipsatively before correlating.

6. Among the more "tentative" methods, which were correlated with the core of "tried" methods on one attitude each (but not with each other), the reliabilities were of the same satisfactory order. Promising validities were found for the methods of Distraction, False Belief and Speed of Decision, suggestions of validity were found for Fluency, while Misperception (Illusion) and Speed of Reading had no validity.

7. All of the tests were very short (10 items each), the purpose of the investigation being only to pick out, among an array of new psychological approaches, those possessed of any validity at all. Lengthening of the tests would raise the validity of six of them to about 0.5, of four others to .3 or .4. Item analysis might raise it somewhat more, but the over-all results seem to indicate that some real specifics are necessarily being measured by the specific methods and that a satisfactory objective measure of an attitude will only be obtained by a battery employing four to six different methods.

8. Various results, notably the existence of a cluster among some methods on the fringe of the main cluster, give slight indications that there is some functional separation of that part of the strength of an attitude which arises from its frustration.

9. From the experience of the four experimenters in the de-

sign and conduct of the experiment some suggestions for improvement when carrying the research further are offered. Together with the methods explored in an extension of this research (8) the present methods constitute a set of *eight new methods* (Information, Immediate Memory, Preference, Speed of Decision, False Belief, Psychogalvanic Response, Projection and Distraction), additional to the criterion methods of Money and Time Expenditure and the classical Opinionaire (which they equal in validity), available for further use. Two directions of research now open up: (a) the improvement of the above valid methods by concentration on each technique singly, in relation to a standard validating core, (b) the exploration of the nine untried methods (Nos. 5, 7, 9, 12, 15, 17, 20, 23 and 24 in the primary list above) described in this same theoretical scheme.

Since the successful contribution of psychology to the much needed integrating studies in the social sciences, with economics, anthropology and sociology, depends to a large extent on the psychologists' ability to supply objective and accurate means of measuring strength of motive, interest or attitude, i.e., of dynamic traits generally, it is to be hoped that the present exploration will be a foundation and stimulus for vigorous research in this area.

The writers wish to express their gratitude to the Graduate Research Board of the University of Illinois and to the Social Science Research Council for funds contributing to the completion of this research.

REFERENCES

1. Bruner, J. S. "Value and Need as Organizing Factors in Perception." *Journal of Abnormal and Social Psychology*, XLII (1947), 33-44.
2. Cattell, R. B. "Experiments on the Psychical Correlate of the Psychogalvanic Reflex." *British Journal of Psychology*, XIX (1929), 357-386.
3. Cattell, R. B. "The Measurement of Interest." *Character and Personality*, IV (1935), 147-169.
4. Cattell, R. B. *The Description and Measurement of Personality*. New York: World Book Company, 1946.
5. Cattell, R. B. "The Ergic Theory of Attitude Measurement." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, VII (1947), 221-246.

6. Cattell, R. B. *A Guide to Mental Testing*. London: Univ. of London Press, 1948.
7. Cattell, R. B. "Principles of Construction of Apperceptive or Projective Tests of Personality." Chapter 2, *Projective Methods* (H. H. Anderson, ed.) New York: Wiley, 1949.
8. Cattell, R. B. "Ergic Structure in Man as Inferred from the Measurement of Attitudes. (In press.)
9. Cattell, R. B., Light, B., Maxwell, F., and Unger, M. "The Objective Measurement of Attitudes." *British Journal of Psychology*. (In press.)
10. Chant, S. N. E. and Salter, M. D. "The Measurement of Attitude Toward War by the Galvanic Skin Reflex." *Journal of Educational Psychology*, XXVIII (1937), 281-289.
11. Colman, R. D. and McCrae, C. R. "An Attempt to Measure the Strength of Instincts." *Education*, V (1927), 171-181.
12. Droba, D. D. "Methods of Measuring Attitudes." *Psychological Bulletin*, XXIX (1932), 309-323.
13. Gallenbeck, C. *Systematic Analysis of the Characteristics of Thinking and Belief*. (In press.)
14. McNemar, Q. "Opinion-Attitude Methodology." *Psychological Bulletin*, XLIII (1946), 289-374.
15. Postman, L. and Zimmerman, C. "Intensity of Attitude as a Determinant of Decision Time." *American Journal of Psychology*, LVIII (1945), 516-518.
16. Smith, W. W. *The Measurement of Emotion*. London: Kegan Paul, 1922.
17. Suppr, D. E. and Roper, E. S. "An Objective Technique for Testing Vocational Interests." *Journal of Applied Psychology*, XXV (1941), 487-498.
18. Thorndike, E. L. "How We Spend Our Time and What We Spend It For." *Science Monthly*, XLIV (1937), 464-469.
19. Thorndike, E. L. "What Do We Spend Our Money For?" *Science Monthly*, XLV (1937), 226-232.
20. Thurstone, L. L. "The Theory of Attitude Measurement." *Psychological Review*, XXXVI (1929), 221-241.
21. Warden, C. J. *Animal Motivation*. New York: Columbia Univ. Press, 1931.

THE CONSTRUCTION AND VALIDATION OF A WORK-TYPE AUDITORY COMPREHENSION READING TEST

GEORGE SPACHE

Chappaqua, New York

WE believe that there is a need for a test to determine the potential ability of students to comprehend and use high-school and college-level reading materials. This test should be relatively free from the influence of intelligence, as commonly measured, and independent of the influence of any reading difficulties of the individual. It should serve to indicate the possible performance level in silent comprehension and auding abilities. In our opinion, such a test would replace the use of common intelligence tests in estimating potential reading ability.

Such a test would be preferable to the use of an intelligence test because the latter is not necessarily a good indicator of potential reading performance. Intelligence is itself a potential which is not achieved to equal degrees in all areas of communication. There is no good reason why an intelligence test should be very closely related to reading ability or more significantly related to comprehension than to writing or speaking skills. We see no reason why one measure of potential general ability should be the best estimate of probable performance in many specific skills.

A second reason against the use of intelligence tests to predict reading comprehension is the extent of common content in such tests. Many intelligence tests actually function as reading tests and their results are merely a measure of reading status rather than an estimate of future or possible performance.

Finally, intelligence tests do not function as accurate measures of potential reading skill because reading performance is not dependent solely upon intelligence. Such factors as exposure to reading materials, socio-economic status, attitudes

toward reading, etc., definitely influence reading performance. These are sufficient to explain many of the observed discrepancies between intelligence and reading test results.

For these reasons, we have attempted to devise a pair of comparable tests that would determine present reading comprehension status and the potential ability of the student to improve his silent comprehension. The tests were arranged to parallel each other by selecting comparable passages from common high school and college texts in science, literature and social science. Two forms of each test comparable in length, difficulty and types of reading passages were constructed. The *Silent Comprehension Test* requires the pupil to read the passages and to answer questions in the usual manner. In the *Auditory Comprehension Test*, passages and questions are read to the student. Thus, we obtain comparable measures of performance and potentiality.

Possible uses of these tests are numerous.¹ The present status of an individual in ordinary silent comprehension can readily be determined. With this knowledge it is possible to detect the extent of comprehension difficulties. The use of the auditory type of test would indicate whether ordinary remedial procedures, or specific training in auding skills (as auditory vocabulary, organizing and summarizing, taking notes, etc.) were necessary or likely to be profitable. To be specific, low scores in silent comprehension in the presence of average or better auditory comprehension would indicate that common remedial techniques would probably be profitable. Low scores in both tests would indicate a degree of low potential for high-school or college work not likely to be improved except by extensive and prolonged remedial help. Average or better scores in silent comprehension with low auditory comprehension would indicate the need for special training in auding or auditory skills.

The results in terms of total score on the first edition of the *Auditory Comprehension Test* were correlated with other sections of the *Diagnostic Reading Test* battery as well as measures of intelligence and reading.

¹ These tests may now be obtained from Dr. Frances O. Triggs, 419 West 119th Street, New York 27, N. Y. They are published by the Committee on Diagnostic Reading Tests, a non-profit corporation devoted to the study and improvement of reading procedures.

In view of a reliability coefficient (S-B) of .788 for this edition, these correlations would seem to support our hope that this test would be a measure of factors operating in reading comprehension. The relationships with the measures of silent comprehension are of the order of 5; those with vocabulary and

TABLE 1
Relations of Scores on Auditory Comprehension Test to Various Other Measures

Auditory Comprehension—Terman McNemar IQ.	358
—Cleveland Reading—Vocab.	358
—Diagnostic Reading—Vocab.	512
Gen Read. Rate.	675
“ “ Comp.	167
Social Studies Rate	493
“ “ Comp.	299
Word Attack.	582
Oral Reading (errors)	400
	177

TABLE 2
Intercorrelation Matrix and Reliabilities (K-R 21) of Total and Part Scores on the Diagnostic Reading Tests, Section II, Comprehension Part 2, Auditory Form A (N = 162)

		Coefficients of Correlation				
		2	3	4	5	6
		K-R 21	Main Ideas (47 Items)	Details (63 Items)	Conclu- sions (25 Items)	Physical Science (47 Items)
						Social Sciences and Literature (88 Items)
K-R 2172	.38	.48	.43	.66
1. Total Score						
No. of Items 13573	.86	.58	.72	.78
2. Main Ideas						
No. of Items 4772		.23	.62	.64
3. Details						
No. of Items 6338		.15	.56	.52
4. Conclusions						
No. of Items 2548			.54	.71
5. Physical Science						
No. of Items 4743				.56

intelligence of the order of 3, with the exception of that with the *Diagnostic Vocabulary Test*; while those with rate, word analysis and intelligence range from 4 downward.

Our finding that there is a significant relationship between silent comprehension and the comprehension of material read to a student is similar to the results obtained by Swanson and,

Anderson.¹ These authors also found that results in the two situations tended to be markedly similar.

A second edition attempted to differentiate questions into subgroups determining the comprehension of main ideas, details and conclusions. Questions on social science and literature were also distinguished from those in physical science in the hope that comprehension in these types of questions and subject matter could be measured. Unfortunately, the intercorrelation matrix of part scores did not support this attempt.

Reliabilities of the subscores range from .38 to .72 and the intercorrelations of sub-sections from .15 to .82. With the possible exception of Main Ideas, none of the subscores is sufficiently reliable to justify its distinction. Item validities ranged in median values from 21.1 for Details, to 40.6 for Main Ideas and 27.1 and 28.5 for Forms A and B, respectively. Median-item validities for Social Science and Literature and for Physical Science questions were 26.4 and 22.7. With this evidence, no attempt was made to differentiate types of questions or subject matter in the final revision.

Before undertaking the third and final revision, we thought it desirable to investigate the influence of chance and informational background upon scores in the *Auditory Comprehension Test*. We have often felt that many questions in other reading tests could be answered by a student without ever reading the test material. In fact, we confirmed this impression in a study of another test in the *Diagnostic Test Battery*. In a measure of silent reading comprehension, we believe that this situation would be highly undesirable, since it would vitiate the attempt to measure comprehension in a specific body of reading materials. Since the purpose of the *Auditory Comprehension Test* is to measure potential, and not performance, in reading, the fact that the student may be able to answer a number of questions even though he has not read the test material does not invalidate the test. If we are to measure potential, then the influence of reading backgrounds and information should be al-

¹ Swanson, D. E., and Anderson, I. H., "A Comparison of Comprehension Scores Obtained from Silent Reading, Oral Reading and Auditory Comprehension." Unpublished research as quoted by D. E. Swanson, in "Common Elements in Silent and Oral Reading," *Psychological Monographs*, XLVIII, (1937) 36-60.

lowed to operate to a reasonable degree since they are contributors to this potential.

A group of 33 high-school pupils whose socio-economic status and intelligence were relatively high were able to answer a median of 58 per cent of the questions correctly. We do not know whether this figure should be greater or less, since we know of no comparable data. It would imply that about half of the questions of the *Auditory Comprehension Test* can be answered on the bases of intelligence, reading background and the other factors that influence reading skills. The remainder of the questions are, presumably, dependent upon the ability to comprehend specific high-school and college textual materials. Thus the test may be measuring potential both by sampling the capacity for understanding a group of selections from common texts and by measuring the facility in using reading or informational experiences.

The third and final editions of the Auditory and the Silent Comprehension tests were based on these experiences with the two preliminary editions. The parallel nature was preserved and the similarity between the tests increased by making them of the same length. The final editions are composed of approximately 50 items and require about one class period for administration. We believe that judicious use of the tests will make possible comparisons between present status and potential performance in reading and auding, as well as a prognosis of the probable outcome of remedial help or training in auding skills.

VALIDATION AND STANDARDIZATION OF THE AGO GENERAL MECHANICAL APTITUDES TEST FOR THE SELECTION OF CIVILIAN EMPLOYEES IN WAR DEPARTMENT INSTALLATIONS¹

ADAM PORUBEN, JR.

Metropolitan Life Insurance Company

DURING World War II, the Civilian Subsection of the Personnel Research Section, The Adjutant General's Office, was engaged in the construction, standardization, and validation of various aptitude tests for the selection and placement of civilian personnel in various War Department installations. The *General Mechanical Aptitudes Test* was one of these tests. It was derived from four tests that already had shown some validity for the selection of employees for mechanical jobs. The study here reported was carried out in 1945. The writer, who was on the staff of the Civilian Subsection, was assigned this particular subject because he had been a teacher of Related Mathematics and Sciences for several years in the Saunders Trades School, Yonkers, N. Y., where this test was tried out, and he was, therefore, in a better position to evaluate the reliability and validity of the criterion data than someone who was not acquainted with the school.

Purpose of Study

The immediate objective of this study was to determine the validity of the *General Mechanical Aptitudes Test* for the pre-

¹ This study was carried out while the writer was on the staff of the Personnel Research Section of The Adjutant General's Office. The opinions expressed in this article are the author's and do not necessarily reflect the official attitude of the Department of the Army.

This article reports only part of this study. The validation study was carried out on six groups of students. This article reports the results obtained on the 11th-year Technical major group. The rest of the study appears in the *Journal of Psychology*, XXIX (1950), 133-155.

The writer makes grateful acknowledgement to Dr. E. E. Cureton and Dr. Erwin K. Taylor for their encouragement in carrying out this study; also to Dr. Lawrence Ashley, Mr. William Carey, and Mr. Patrick McLugh, who permitted this study to be carried out in the Saunders Trades School, Yonkers, N. Y.

diction of success in industrial and technical high schools. The ultimate aim was to determine its validity for the selection of civilian employees for various mechanical jobs in War Department Installations. Since most of the graduates of the Saunders Trades School go into their respective trades and specialties upon graduation and are fairly successful in their work, it was hoped that the validation of the *General Mechanical Aptitudes Test* for the prediction of success in this school would also give some indication of its validity for the selection of employees for various mechanical jobs which are similar to those for which the students were being trained in this particular school.

The School

The Saunders Trades School is an industrial and technical senior high school for boys, supported mostly by Federal, State, and local funds, but partly by private funds derived from the so-called Saunders Fund. This school serves the entire city of Yonkers, N. Y.; its normal enrollment is over 1,000 students. At the time of this study, however, the enrollment was considerably under this figure because of war conditions.

The Saunders Trades School offers two majors, industrial and technical. The industrial major has a duration of three years, the students being admitted after they have completed the ninth grade in one of the several junior high schools in Yonkers. This major is more or less terminal in nature in that most of the boys are expected to go to work in their respective trades upon graduation. The technical major also has a duration of three years. Its graduates are expected either to become junior engineers in industry or to go to engineering colleges for further study. The series of courses in this major are so arranged that the students can meet college entrance requirements upon graduation.

The industrial major consists of seven curricula: Auto Mechanics, Building Maintenance, Carpentry, Electric Installation, Machine Shop, Plumbing, and Refrigeration. A student entering the industrial major, selects one of these curricula. All seven curricula are parallel in nature, but in each one the student pursues course and shop work along his particular

specialty. The student devotes about half of his time to the shop and laboratory courses and the other half to related courses in mathematics, science and drafting. The work in these courses is fairly well integrated with the work in the shops, that is, the theory and the mathematics involved in a particular shop instruction unit are first discussed in the related courses before the shop work is begun. For example, the student is taught the theory and mathematics of parallel circuits in electricity before he performs the experiment in that project in the shop. The degree of integration between the related courses and the shop projects varies among the seven curricula, depending largely on the cooperation of the instructors.

The technical major consists of five curricula: Architecture, Industrial Chemistry, Electricity, Machine Design, and Power Generation. Each student pursues one of these for three years. As in the industrial major, the five curricula are parallel in nature, but, at the same time, specialized. Each curriculum consists of shop or laboratory work, related courses, and certain academic courses such as English and American History. The related courses are integrated with the shop and laboratory work.

The Population

This study was carried out on one of the three groups of students in the technical major, namely, the 11th-year group. There were seventy-two students in this group distributed among the five curricula as follows:

<i>Curricula</i>	<i>Number of Students</i>
Architectural Course	10
Industrial Chem. Course	8
Electrical Course	18
Machine Design Course	22
Power Generation Course	14
Total	72

Description of the Test

The *General Mechanical Aptitudes Test* was designed to measure various aspects of mechanical aptitude. It consists of four subtests as follows:

1. Mechanical Comprehension.—This consists of 43 three alternative multiple-choice items administered with a 15-minute time limit. The items of this test were adapted, by permission of the author, from Forms AA, BB, and WI of the *Bennett Test of Mechanical Comprehension*. It measures general mechanical insight and the capacity of an individual to understand mechanical operations.

2. Technical Reading.—This is a paragraph-and-question test based on selections from technical manuals and texts. It consists of 29 items administered with a 15-minute time limit. The directions and a sample question are shown below.

DIRECTIONS

This test consists of five paragraphs and some questions about each paragraph. There are 29 questions in all. Read each paragraph and then answer the questions which follow. Read the paragraph as many times as you need to in order to answer the questions. The first paragraph and the questions based on it is a sample to show you what to do.

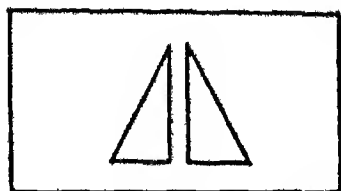
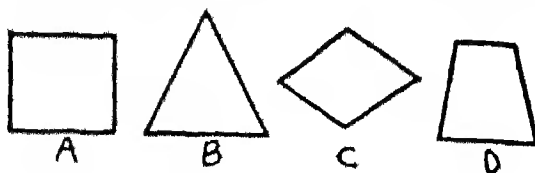
The blast furnace is a great stone chimney 100 feet high or more. It is filled with a roaring fire from top to bottom. Into the top of the blast furnace are dumped carefully measured amounts of iron ore, coke, and limestone. After 4 or 5 hours of terrific heat, molten iron is drained off from a door at the bottom of the furnace.

46. A blast furnace is made of
A. iron
B. stone
C. clay
D. coke

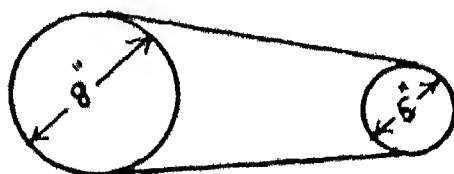
3. Paper Form Board.—This consists of 44 items, administered with a 10-minute time limit, and measures the ability to manipulate spatial images mentally. The directions for the test and a sample question follow.

DIRECTIONS

This test consists of 45 problems. At the top of each page, there are four large figures labeled A, B, C, and D, like those shown in the first row below. Each problem shows one of these figures cut into pieces and scattered around in a box. Look at the pieces in each box, and decide which one of the figures could be made if all the pieces in that box were fitted together. Some of the pieces may need to be turned around or turned over to make them fit. The pieces in each problem will make only one of the figures. The first problem is a sample.



4. Shop Arithmetic. This test consists of 20 free-answer arithmetic reasoning problems based on shop arithmetic. Sixteen of the items contain diagrams, tables, or drawings, and the test is administered with a 20-minute time limit. A sample problem is shown below:



When the larger pulley wheel makes 100 turns per minute, how many turns per minute does the smaller wheel make?

Procedure

Before the *General Mechanical Aptitudes Test* was given, permission was obtained from the Yonkers school authorities to transcribe the school grades for use as criteria. These grades were copied from the school's progress sheets which list the grades according to the curriculum and the year. In May, 1945, the test was administered to 480 students of the Saunders Trades School by the teachers after a training session had been given by a staff member of the Personnel Research Section. All tests were scored at the headquarters of the Personnel Research Section.

Analysis and Results

In order to see whether the seventy-two students used in this study constituted a fairly homogeneous group, the analysis of variance technique was used to investigate the question whether the four subtests of the *General Mechanical Aptitudes Test* differentiated significantly among the five curricula of the technical major. The results, which are not reported here, showed no significant F-ratios at the 1 per cent level among the curricula within the technical major. Therefore, the 11th-year students from the five technical major curricula were combined into one group and the analysis carried out on this group.

1. *The Criterion.*—In a technical high school, such as the Saunders Trades School, each technical subject has a definite place in the total pattern of instruction offered; that is, the class-room subjects such as mathematics, science, and theory, are definitely related to the shop or laboratory work. These class-room subjects provide the student with the basic knowledge and fundamental skill which will enable him to pursue his shop studies more intelligently. For example, in the Applied Mathematics course the students in the Electrical curriculum learn the basic mathematics connected with the series and parallel circuits; in the Basic Theory course they study the fundamentals of the series and the parallel circuits. With this background, the student can learn the shop work more easily and more intelligently. Moreover, the work in the classroom is fairly well integrated with the work in the shop; that is, the theory and the mathematics involved in a particular shop instruction unit are first discussed in the related courses before the shop work is begun.

Because of this high integration of related subjects with the shop work, and because the technical subjects do represent the core of the curriculum of the Saunders Trades School, it occurred to the writer that a composite of the grades in these technical subjects would constitute a more valid criterion than a composite of all of the grades, including the more academic subjects such as English, Economics, History, etc. Because of these considerations, the composite of the grades received by the students during their 10th and 11th years in

the technical subjects was taken as the criterion in the study. This composite was obtained by the summation of 12 grades in five different subjects, namely, Basic Theory, Shop, Physics, Applied Mathematics, and Plane Geometry.

Estimate of the reliability of the criterion was obtained by correlating the sum of scores for the first terms with the sum of scores for the second terms. This reliability coefficient was found to be .88. When this was stepped up by the Spearman-Brown formula, it became .94.

TABLE 1
Intercorrelations Among Tests and the Criterion

	(N = 72)				
	Mech. Comp.	Tech. Reading	Paper Form Board	Shop Arith.	Criterion
Mean	19.556	17.667	33.972	13.653	919.940
S.D.	7.708	5.664	5.475	2.800	83.650
Mech. Comp.		.531	.476	.394	.493
Tech. Reading			.217*	.563	.542
Paper Form Board				.262*	.417
Shop Arithmetic					.454

* Not significant at the one per cent level

2. *Reliabilities of the Tests.*—No estimates of reliabilities of the four tests were made in this study. Such estimates, however, were made previously by the Personnel Research Section, and were found to be quite satisfactory.

3. *Intercorrelations.*—The intercorrelations among the tests and the criterion are shown in Table 1. All of the correlations, except two, were found to be significantly different from zero at the one per cent level.

4. *Multiple Correlations.*—The multiple correlation was computed by the usual Doolittle method and also by the Wherry-Doolittle method in order to show the amount of shrinkage in the R. By the Doolittle method, the multiple R for the entire battery was found to be .644 and .624 by the Wherry-Doolittle method. Thus, the shrinkage in the multiple R was fairly small, namely, .02.

When the shrunken multiple R was corrected for the attenuation in the criterion, it became .643 or $R^2 = .413$. Thus the battery accounts for about 41 per cent of the variance of the criterion.

In order to show the contribution of each test toward the efficiency of the battery, Table 2 is presented.

TABLE 2
Contributions of the Tests Toward Battery Efficiency

Test Battery	\bar{R}^2	\bar{R}	Corrected* \bar{R}	% of Criterion Variance
B	.2938	.542	.559	31.2
B, C	.3790	.616	.635	40.3
B, C, D	.3870	.622	.641	41.1
B, C, D, A	.3891	.624	.643	41.3

A = Mechanical Comprehension

B = Technical Reading

C = Paper Form Board

D = Shop Arithmetic

* Corrected for the attenuation in the criterion, the reliability coefficient being .94.

5. *Beta Weights*.—The beta weights were also found by the Doolittle as well as the Wherry-Doolittle methods. Their values are shown below:

Test	Beta Weights
Technical Reading328
Paper Form Board240
Shop Arithmetic153
Mech. Comprehension.137

6. *Regression Equation*.—The regression equation for predicting the criterion from the four subtests of the *General Mechanical Aptitudes Test* may be written in standard form as follows:

$$\bar{z}_c = .328 z_B + .240 z_O + .153 z_D + .137 z_A$$

In order to get the equation in score form, the β 's were transformed into the corresponding b 's. The resultant equation, expressed in terms of deviations from the mean, is as follows:

$$\bar{x}_c = 4.844 x_B + 3.667 x_O + 4.571 x_D + 1.487 x_A$$

with a standard error of estimate of 64.08.

Conclusions

1. In general, the *General Mechanical Aptitudes Test* shows fairly high validity for the prediction of academic success in the basic technical courses in an industrial or technical high school. The multiple correlation, when corrected for attenuation, was found to be .643. Thus, the *General Mechanical Aptitudes Test* battery accounts for about 41 per cent of the variance of the criterion.

When the Mechanical Comprehension Subtest is removed from the battery, the multiple R, corrected for attenuation, is .641 and the resultant three-test battery still accounts for about 41 per cent of the variance of the criterion. Thus, for a forty five minute testing time, one can get a fairly good indication of the probable success of a student in a technical high school such as Saunders.

In the larger study, the efficiency of the battery for the prediction of success in specific subjects such as mathematics, science, shop, and theory, was studied. None of the multiple R's in that study exceeded the multiple R found here. Only one of them, the one predicting success in Physics, equaled the multiple R found in this study.

2. From the Beta Weights one can conclude that the *Technical Reading Test* contributes most heavily to the prediction efficiency of the battery. There is so much technical reading required in the science, mathematics, theory, and shop courses of the technical high school that skill and speed in doing such reading contributes heavily towards academic success in such a school.

3. Saunders Trades School is generally similar in students, curriculum, instructors, support, etc., to other industrial and technical high schools in New York State. Therefore, one can conclude that the *General Mechanical Aptitudes Test* is a valid test for the prediction of success in an industrial and technical high school in the state of New York. Since many states follow the N. Y. State pattern of vocational education, one could probably conclude that this test has similar validity for any such industrial or technical high school.

4. Finally, since most of the graduates of the Saunders Trades School go into their respective trades and specialties upon graduation and are fairly successful in their work, the writer would conclude that this test should be fairly valid for the selection of employees for the various mechanical jobs for which training is given in this school. In other words, from this study one can infer that the *General Mechanical Aptitudes Test* should be valid for the selection of machinists, machine designers, electricians, power plant operators and technicians, junior industrial chemists, and junior architects.

THREE AIDS IN THE EVALUATION OF THE SIGNIFICANCE OF THE DIFFERENCE BETWEEN PERCENTAGES

C H LAWSHE and P. C. BAKER

Purdue University

THOSE who construct and use paper-and-pencil tests are confronted with the task of making a lengthy item analysis. Many authors have offered various devices to aid in this work. All of these serve the purpose for which they were intended; however, there are within and among them several weaknesses, viz:

1. They may not be truly time-saving. A considerable amount of additional computation may be required.
2. They may give only gross approximations to the desired values.
3. They may give results in terms of a statistic whose sampling error distribution is unwieldy.

We here offer three instruments to be used in the evaluation of the significance of the difference between two percentages. Table 1, "The Significance of the Difference Between Percentages"; Table 2, "The Omega Equivalent to a Percentage"; Figure I, a nomograph to estimate the significance of the difference between percentages.

These instruments were devised with certain criteria in mind:

1. The amount of calculation required shall be minimal
2. Restrictions placed upon the data shall be minimal
3. The results obtained shall be accurate to a degree commensurate with published results
4. The results obtained shall be subject to a "standardized" interpretation. There shall be no ambiguity.
5. The instruments shall be compact; easy to use.

Table 1.—Table 1 is the result of a direct approach to the usual formula for the critical ratio of the difference between two percentages to the standard error of that difference when

TABLE I

The Significance of the Difference Between Percentages

$\frac{N_1}{N_2}$	100	95	90	85	80	75	70	65	60	55	50	45	40	35	30	25	20	15	10	5
0			4.3589	3.0000	2.3805	1.6000	1.7301	1.5375	1.3028	1.1243	1.0000	.8845	.8167	.7508	.6847	.6174	.5500	.4817	.4124	.3424
5	4.3589	2.9200	2.2923	1.9124	1.6405	1.4445	1.2809	1.1442	1.0218	.9204	.8311	.7508	.6847	.6174	.5500	.4817	.4124	.3424	.2724	.2024
10	3.0000	2.5923	1.8836	1.6031	1.4000	1.2359	1.0954	.9781	.8702	.7734	.6869	.6023	.5212	.4424	.3657	.2911	.2174	.1447	.0724	
15	2.3805	1.9124	1.6032	1.3862	1.2123	1.0490	.9467	.8392	.7413	.6512	.5682	.4870	.4074	.3291	.2519	.1757	.1004	.0261		
20	2.0000	1.6405	1.4000	1.2123	1.0667	.9330	.8225	.7259	.6335	.5483	.4687	.3916	.3161	.2424	.1704	.1000	.0311			
25	1.7301	1.4445	1.2359	1.0660	.9330	.8165	.7137	.6259	.5452	.4687	.3924	.3174	.2436	.1709	.1000	.0311				
30	1.5375	1.2809	1.0954	.9467	.8225	.7137	.6172	.5292	.4472	.3696	.2909	.2171	.1428	.0710						
35	1.3028	1.1442	.9781	.8392	.7259	.6259	.5392	.4472	.3696	.2909	.2171	.1428	.0710							
40	1.2809	1.0218	.8702	.7413	.6512	.5682	.4870	.4074	.3291	.2519	.1757	.1000	.0311							
45	1.1055	.9205	.7746	.6532	.5483	.4549	.3696	.2909	.2171	.1428	.0710									
50	1.0000	.8251	.6860	.5696	.4685	.3780	.2940	.2171	.1428	.0710										
55	.9045	.7305	.6025	.4809	.3916	.3032	.2188	.1431	.0710											
60	.8165	.6508	.5222	.4124	.3162	.2294	.1491	.0731												
65	.7338	.5721	.4437	.3357	.2410	.1552	.0793													
70	.6547	.4927	.3651	.2582	.1644	.0793														
75	.5774	.4126	.2847	.1782	.0848															
80	.5000	.3393	.2000	.0933																
85	.4201	.2590	.1072																	
90	.3333	.1748																		
95	.2500																			

the size of the samples upon which the two percentages are based are equal. (It is applicable only when $N_1 = N_2$.)

$$t = \frac{p_1 - p_2}{\sqrt{\frac{p_1 q_1}{N} + \frac{p_2 q_2}{N}}} \quad (1)$$

$$\frac{t}{\sqrt{N}} = \theta = \frac{p_1 - p_2}{\sqrt{p_1 q_1 + p_2 q_2}} \quad (2)$$

Let the right-hand member of equation 2 equal Theta. Theta could be evaluated for all combinations of p_1 and p_2 , but this

TABLE 2
The Omega Equivalent to a Percentage
(Omega positive for $p > .50$, negative for $p < .50$)

p	Ω	p	p	Ω	p
1.00	1.1106	.00	.75	.3702	.25
.99	.9690	.01	.74	.3541	.26
.98	.9099	.02	.73	.3379	.27
.97	.8648	.03	.72	.3221	.28
.96	.8259	.04	.71	.3064	.29
.95	.7917	.05	.70	.2910	.30
.94	.7602	.06	.69	.2756	.31
.93	.7320	.07	.68	.2612	.32
.92	.7052	.08	.67	.2453	.33
.91	.6797	.09	.66	.2303	.34
.90	.6565	.10	.65	.2155	.35
.89	.6326	.11	.64	.2006	.36
.88	.6104	.12	.63	.1861	.37
.87	.5891	.13	.62	.1714	.38
.86	.5697	.14	.61	.1568	.39
.85	.5472	.15	.60	.1424	.40
.84	.5287	.16	.59	.1280	.41
.83	.5096	.17	.58	.1137	.42
.82	.4911	.18	.57	.0994	.43
.81	.4728	.19	.56	.0851	.44
.80	.4550	.20	.55	.0708	.45
.79	.4375	.21	.54	.0567	.46
.78	.4202	.22	.53	.0424	.47
.77	.4033	.23	.52	.0283	.48
.76	.3867	.24	.51	.0141	.49
.75	.3702	.25	.50	.0000	.50

would result in a table much too large for practical use; hence, we have limited ourselves in Table 1 to combinations of p_1 and p_2 which are multiples of five.

To use Table 1 it is necessary only to multiply the tabled value of Theta by the square root of N to find the critical ratio.

$$t = \theta \sqrt{N} \quad (3)$$

Table 1 is useful in classifying a large number of differences into three categories; (1) definitely significant, (2) doubtful,

(3) definitely not significant. Those differences falling in the doubtful category may then be more carefully evaluated by means of equation 1.

Table 2. Further consideration of the inaccuracies inherent in Table 1 due to the skewness of the sampling distribution of p when the true value of p approaches 100% or 00% led to the development of a statistic which is a function of p and which has a constant standard error dependent only on the size of the sample. Readers familiar with Fisher's arctanh transformation of r will recognize the utility of such a statistic. Kelly (9, pp. 593-594) offers the development of such a statistic; our development differs only in minor details.

$$\Omega = f(p) \quad (\text{Omega is a function of } p)$$

$$d\Omega = f'(p)dp \quad (\text{First derivative})$$

$$\sigma_{\Omega}^2 = [f'(p)]^2 \sigma_p^2 \quad (\text{Take variance of both sides})$$

$$\sigma_{\Omega}^2 = \frac{1}{N} \quad (\text{Let variance error be inversely proportional to } N)$$

$$\frac{1}{\sqrt{N}} = f'(p) \sqrt{\frac{pq}{N}}$$

$$\frac{1}{\sqrt{p} - p^2} = f'(p)$$

$$2 \int \frac{1}{\sqrt{1-p}} dp = \int f'(p) \quad (\text{Integrate})$$

$$f(p) = 2 \arcsin \sqrt{p} + C$$

$$\Omega = 2 \arcsin \sqrt{p} + C \quad (\text{The desired function})$$

To test the significance of the difference between two percentages we need only to transform them to Omegas and apply the critical ratio formula:

$$t = \frac{\text{difference}}{\text{S. E. of difference}}$$

$$t = \frac{(2 \arcsin \sqrt{p_1} + C) - (2 \arcsin \sqrt{p_2} + C)}{\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

$$t = \sqrt{\frac{2N_1 N_2}{N_1 + N_2}} \left[\sqrt{2} \left(\arcsin \sqrt{p_1} - \frac{\pi}{4} \right) - \sqrt{2} \left(\arcsin \sqrt{p_2} - \frac{\pi}{4} \right) \right] \quad (4)$$

The reason for the algebraic manipulation, factoring out the square-root of two in the numerator, will be clear from the

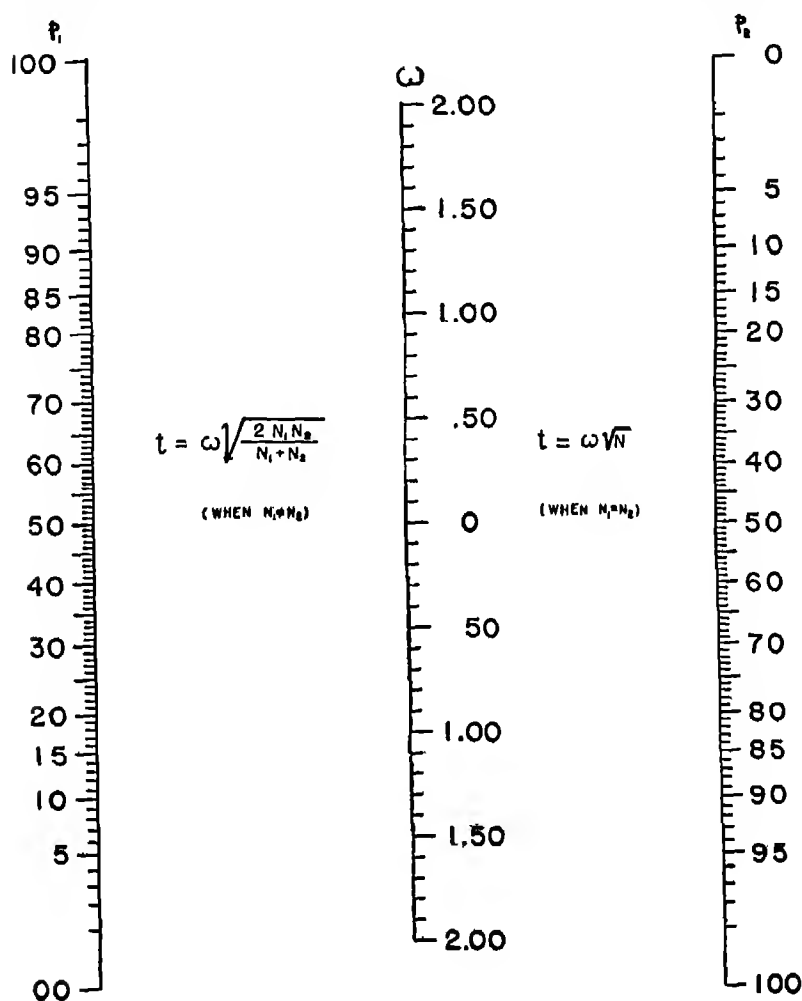


FIG. I.

following. The constant of integration is evaluated as $\sqrt{2} \frac{\pi}{4}$ in order to make the function symmetrical about 50%.

If N_1 equals N_2 formula 4 reduces to

$$t = \sqrt{N} (\Omega_1 - \Omega_2)$$

Omega is now defined as

$$\Omega = \sqrt{2} \left(\arcsin \sqrt{p} - \frac{\pi}{4} \right)$$

Table 2 contains values of Omega for all values of p . These values are positive for p greater than 50% and negative for p less than 50%.

For simplicity in notation we here define Omega lower case as the difference between two Omegas

$$\omega = \Omega_1 - \Omega_2$$

TABLE 3
The 1%, 5%, and 10% Confidence Values of Theta and Omega

Confidence Level	Use with Table 1 (N_1 & N_2 are equal)	Use with Table 2 and with Figure 1 (N_1 & N_2 are equal) (N_1 & N_2 are unequal)
1%	$2.5758 = \theta_{.01}$ \sqrt{N}	$2.5758 = \omega_{.01}$ $\sqrt{\frac{2N_1N_2}{N_1 + N_2}}$
5%	$1.9600 = \theta_{.05}$ \sqrt{N}	$1.9600 = \omega_{.05}$ $\sqrt{\frac{2N_1N_2}{N_1 + N_2}}$
10%	$1.6449 = \theta_{.10}$ \sqrt{N}	$1.6449 = \omega_{.10}$ $\sqrt{\frac{2N_1N_2}{N_1 + N_2}}$

To use Table 2 find the Omega equivalents of the two percentages, find the algebraic difference between the two Omegas, multiply this difference by $\sqrt{\frac{2N_1N_2}{N_1 + N_2}}$ if the two samples differ in size, or multiply the difference by \sqrt{N} if the samples are of equal size. This yields a critical ratio which can be evaluated in terms of the normal probability function, or Student's distribution of t if $N_1 + N_2$ is less than 30, in which case the degree of freedom associated with t is equal to $N_1 + N_2 - 2$.

$$t = \omega \sqrt{\frac{2N_1N_2}{N_1 + N_2}}$$

$$t = \omega \sqrt{N}$$

Figure I.—Figure I is a graphic representation of Table 2. To use this nomograph, find p_1 and p_2 , and join these points

by a straightedge. Where the straightedge crosses the center scale find omega (ω). This value is identical with that found from Table 2, i e., $\omega = \Omega_1 - \Omega_2$, and is used in the same manner. This nomograph¹ is of use when a large number of differences are to be evaluated and classified as significant, doubtful, and not significant.

A Shortcut—When a great many differences are to be evaluated, as in an item-analysis study, the following shortcut is suggested. Instead of multiplying each Theta or Omega value by the square-root of N, find the Theta value or Omega value corresponding to critical ratio values significant at various desired levels of confidence by performing the operations suggested in Table 3, "The 1%, 5%, and 10% confidence values of Theta and Omega."

Figure I can be converted into a "tailor-made" nomograph for one particular study by marking the 1%, 5% and 10% confidence values of Omega on the center scale, and by writing frequency values corresponding to proportions of N_1 and N_2 along the p_1 and p_2 scales.

Summary

Three instruments, two tables and a nomograph, to be used in the evaluation of the significance of the difference between two percentages have been offered. The research worker with a large number of differences to evaluate can, with three simple calculations, determine which of his items attain the 1%, 5%, or 10% levels of significance.

REFERENCES

1. Burr, I. W. and Hobson, R. L. "Significance of Differences in Proportions with Constant Sample Frequencies in Each Pair." *Journal of Educational Research*, XXXIV (1943), 307-312.
2. Davis, F. B. *Item Analysis Data*. Harvard Education Papers, No. 2. Cambridge, Mass.: Harvard Univ. Press, 1946.
3. Dunlap, J. W. and Kurtz, A. K. *Handbook of Statistical Nomographs, Tables, and Formulas*. Yonkers-on-the-Hudson: World Book Company, 1932.
4. Dunlap, J. W. "Note on Computation of Bi-Serial Correlations." *Psychometrika*, I(1936), 51-58.

¹ The authors have copies of the nomograph (8½" x 11") which they will supply gratis in response to single copy requests. Address the senior author.

5. Dunlap, J. W. "Nomograph for Computing Bi-Serial Correlations." *Psychometrika*, I(1936), 59-60.
6. Edgerton, H. A. and Paterson, D. G. "Table of Standard Errors and Probable Errors of Percentages for Varying Numbers of Cases." *Journal of Applied Psychology*, X(1926), 378-391.
7. Guilford, J. P. "The Phi Coefficient and Chi Square as Indices of Item Validity." *Psychometrika*, VI(1941), 11-19.
8. Jurgensen, C. E. "Tables for Determining Phi Coefficient." *Psychometrika*, XII(1947), 17-29.
9. Kelly, T. L. *Fundamentals of Statistics*. Cambridge, Mass: Harvard Univ. Press, 1947.
10. Lawshe, C. H., Jr. "A Nomograph for Estimating the Validity of Test Items." *Journal of Applied Psychology*, XXVI(1942), 846-849.
11. Lichte, W. H. "A Method and Tables for Obtaining Standard Errors of Differences Between Proportions When N is Equal to N." *Journal of Applied Psychology*, XXXI(1947), 449-456.
12. Long, J. A. and Sandilford, P. *The Validation of Test Items*. Department of Educational Research, No. 3. Toronto: University of Toronto, 1935.
13. Moster, C. I. and McQuitty, J. V. "Methods of Item Validation and Abacs for Item Test Correlations and Critical Ratio of Upper-Lower Difference." *Psychometrika*, V(1940), 57-65.

A STUDY OF FAKING ON THE KUDER PREFERENCE RECORD¹

ORRIN H. CROSS
University of Alabama

SINCE vocational guidance counselors and employment offices of industrial concerns use it so frequently, the author of the present paper felt that the possibility of faking the *Kuder Preference Record* needed investigation. Examination of the items of this inventory reveal many which apparently would be quite transparent even to the average individual. This possibility throws some doubt on the advisability of using it except when wholehearted cooperation of the subject is assured.

A recent paper by Longstaff (1) on a similar problem has indicated that both the Kuder and the *Strong Vocational Interest Blank for Men* are susceptible to multiple faking, i. e., faking upward on some of the scales and downward on the remaining ones. The present study differs from Longstaff's in several ways. In the first place, Longstaff's subjects were mature students in an evening Extension Division class in Vocational Development and Personnel Psychology at the University of Minnesota, presumably somewhat sophisticated in psychological test taking; the subjects for this study were drawn from a high-school group, probably quite unsophisticated psychologically. Secondly, Longstaff had his subjects attempt multiple faking, upward on the mechanical, scientific, artistic, literary, and musical scales of the Kuder, downward on the remainder; the subjects for the present study were instructed to fake either up or down on just one scale at a time. Finally, check studies were made in the present case to determine whether previous acquaintance with the test might have been

¹ This research was supported in full by a grant from the Research Committee of the University of Alabama. Papers based on part of these data were presented at the 1949 meetings of the APA and the Southern Society for Philosophy and Psychology.

a factor in success and also whether differences in age and education might have been a factor.

Procedure

The construction and standardization of the test is reported elsewhere (2) and consequently will not be reviewed here.

Two short preliminary studies were done in a small southern high school, both of them on one scale (the Mechanical), with one sex (male). In the first study, all seventh- and eighth-semester students who could find time to take the inventory were tested. The highest ten males on the Mechanical scale were then asked to re-take the test with the instructions to fake a low interest in the mechanical field of work. Several of the lowest ten also re-took it with instructions to fake high interest. Both groups were successful.

In the second preliminary study, the procedure was varied in that the students were asked to fake a high mechanical interest prior to any acquaintanceship with the test. The difference obtained between this group of 36 boys and that of the Kuder norm group of high-school boys proved to be significant at the .01 level (actually there were fewer than 6 chances in 100,000 that such a difference could occur by chance).

The study being reported here used the method of the first preliminary study,² i. e., (1) honest test administered by school authorities, (2) selection of high and low scoring individuals on each scale, (3) retest with instructions to "fake" an interest in the opposite direction.³ The subjects had not been informed of the results of the honest test.

² In order to secure the cooperation of the high-school authorities, this procedure had to be followed.

³ A copy of the directions to fake follow:

Directions

The inventory you did previously was done as a student; a non-employed student honestly giving a picture of his own interests. We would now like your cooperation in doing this inventory as a person who was looking for a special kind of job might; a person who might want to "fool" the test. We want you to help us find out if this can be done.

Directions for Faking Low

You are now to pretend that your doctor has warned you that to take——*work would mean almost certain death. If you show high interest in that kind of work you will be forced to take it in spite of this fact.

You must "slant" the test results so that you will not have to take this type of work; so that you show as little interest in——*work as you can. Thus you will score the one

About six hundred (600) high-school students in four of the five high schools of a large southern city took the honest test. From this group were selected 364 students for retesting—181 males and 183 females. It seemed desirable to compare within the sexes because norms on the sexes differ.

A check study with college students from beginning psychology classes was made, using the method of the second preliminary study, i. e., fake test without previous experience with the inventory. Means and standard deviations were computed for a comparable college group from the same campus for an honest taking of the inventory. Comparisons were then made between the two college groups. This group was not asked to fake low because it appeared to the author that faking low would not be a significant problem in the situations in which the results of the research would be useful. In the guidance and industrial situations the peaks of a profile are regarded as of positive significance, while the low scores are commonly used for their negative value, if at all.

Finally, a group of 67 college students who had taken the inventory were asked to rank the interest fields in order of what they thought their test profiles would show. In this part of the study, a list of the scales was presented, each one followed by a list of from four to fifteen of the occupations listed by Kuder in his Manual as being representative of the occupations in that field of major interest.

Results and Discussion

It will be noted from Table 1 that the high-school students were quite successful at the assigned task, the probabilities

item of each trio which appears to you to be *least* indicative of—* interest in column headed "1", and the one *most* indicative of—* interest in column "3", make no mark opposite the other activity

Directions for Faking High

You are to pretend you very much want a particular job. If you show a large amount of—* interest on this test you have it "cinched". The job does not necessarily involve—* work, but you must show very high *interest* in such things.

You must "slant" the test results so that you will appear to have a great deal of interest in—* work. Thus you will score the one item in each trio which appears to you to be *most* indicative of—* interest in the column headed "1"; and the one *least* indicative of—* interest in the column headed "3"; make no mark opposite the other activity.

*The name of the scale being faked was inserted in each of the blanks. At the end of the directions was appended a list of the occupations chosen from Kuder's lists for the scale being faked.

TABLE 1
Fake versus Honest Test of High-School Students

Scale	Males						Females					
	Fake High			Fake Low			Fake High			Fake Low		
	df	"t"	P	df	"t"	P	df	"t"	P	df	"t"	P
1. Mechanical.....	7	7.445	<.01	9	49.4	<.01	9	37.67	<.01	9	17.11	<.01
2. Computational.....	10	23.78	<.01	13	5.31	<.01	10	16.11	<.01	9	3.89	<.01
3. Scientific.....	10	9.73	<.01	9	3.00	<.01	8	15.924	<.01	9	3.63	<.01
4. Persuasive.....	7	17.23	<.01	9	12.87	<.01	13	8.638	<.01	8*	3.23	.02 > P > .01
5. Artistic.....	11	9.33	<.01	8	7.67	<.01	9	16.56	<.01	12	7.63	<.01
6. Literary.....	10	35.12	<.01	8	13.43	<.01	9	12.25	<.01	8	6.92	<.01
7. Musical.....	11	22.96	<.01	7	7.59	<.01	10	14.92	<.01	12	12.47	<.01
8. Social Service.....	10	23.24	<.01	10	4.63	<.01	9	9.108	<.01	8	25.86	<.01
9. Clerical.....	9	8.22	<.01	10	5.61	<.01	5	9.629	<.01	11	21.23	<.01

* 3.355 = .01 Corrected for 1 deviant case $t = 6.999$ $P < .01$

that the observed differences were due to chance being less than .01, with the exception of the females faking low on the Persuasive scale. On this scale, the probabilities lay between .02 and .01. If the results are corrected for a deviant case⁴, who actually raised her score, this probability also drops to less than .01. Each sex was compared to the Kuder Manual norms as a measure of its faking ability. Neither sex failed to fake high successfully on any one of the scales. On the other hand, both sexes failed to successfully fake low on scales four (Persuasive) and nine (Clerical), and the females also failed on scale two (Computational). If these results are each corrected for a single deviant case whose fake score proved to be *higher* rather than lower on the retest, the probabilities drop to less than .01 on all the scales except the Persuasive for females. This probability is .075. Comparison of the sexes, scale by scale, failed to reveal any significant differences between them. The "t" values ranged between .090 and .681 for from 13 to 21 degrees of freedom.

College males and college females proved about equally excellent at faking high when compared to the norms derived in this study. For the male group the "t" values ranged between 3.93 and 25.41 (median = 12.09) for degrees of freedom between 63 and 66 for the various scales. For the female group the "t" values fell between 6.49 and 26.88 (median = 10.04) for degrees of freedom between 128 and 134. Comparison of the sexes revealed significant differences between them on the Persuasive scale only, with the males showing superiority there. The "t" value here was 6.353 for 21 degrees of freedom.

Finally, the high-school and college groups were compared. No significant differences were evident here. The "t" values obtained fell between 2.04 (for 14 degrees of freedom) and .049 for from 14 to 21 degrees of freedom.

Inspection of the pertinent data (Table 1) indicates that faking high is easier than faking low. The male sex faked high better than low on all but the Mechanical scale, the female sex

⁴ The author assumed that when a subject instructed to fake low not only failed to lower his score on the retest, but *raised* it, he was not following directions either because he misunderstood or because he did not wish to cooperate. Critical ratios, except where noted, were calculated with such deviant cases included.

faked high better than low on all but the Scientific, Social Service, and Clerical scales. Faking high is a more important ability in the situations in which the test is applied, consequently this finding appears pertinent.

Faking high appears to be somewhat easier for the college group than for the high school group although the difference does not reach the .01 level of significance on any scale for either sex. Differences favoring the high-school groups occurred on the Computational and Artistic scales for the male sex, and for the Persuasive and Artistic scales for the female sex.

The mean rank difference correlation of the college group which attempted to predict what its order of standing on the nine scales would be was $+.67$. This coefficient is significant at the .01 level.

The uses to which such an inventory is put need examination. First, it is used in vocational and educational guidance in public schools, colleges, guidance centers, and the employment services; and, second, it is used in the selection of workers for a job in business and industry.

What do the results reported mean, then? In the first case, guidance, rapport may be justifiably assumed. In this case, such a fakable inventory retains usefulness. However, in the second instance no such assurance of cooperation exists. On the contrary, it might reasonably be assumed that testees take the opposite attitude, consciously or unconsciously. On the basis of the reported results, it might be asserted that such an inventory as this must be interpreted with caution in the industrial situation, at least where the applicant has any inkling of the job he is being considered for.

Longstaff (1) has suggested the analysis of this inventory after the fashion of the "K scale" of the *Minnesota Multiphasic Personality Inventory* to seek a correction for faking. That was the original intent of the study here reported, but analysis of some of the data obtained failed to reveal enough items for such a scale. Another possibility in selecting workers for a job might be the use of the whole profile, on the assumption that secondary (and less transparent) peaks, and low scores might prove to be discriminating.

Conclusions

The results reported above appear quite consistent in their implications. In only one case did a group (in this study, the high-school females, faking low on the Persuasive scale) fail to perform the task successfully as compared to its own honest tests. Correction of this result for a deviant case brought that result to significance also. It thus appears that a subject suitably motivated may successfully fake the *Kuder Preference Record*

As shown by the present study, when an applicant for a job has any idea of what job he is being considered for, his scores should be interpreted in the light of the knowledge that faking is possible *if he desires to fake*. In the properly motivated guidance situation, this problem does not arise.

REFERENCES

- 1 Longstaff, H. P. "Fakability of the Strong Interest Blank and the Kuder Preference Record." *Journal of Applied Psychology*, XXXII (1948), 360-369
2. *Revised Manual for the Kuder Preference Record* Chicago: Science Research Associates, 1946.

PSYCHOLOGICAL TESTING FOR IMMIGRANTS IN A VOCATIONAL COUNSELING AGENCY¹

BENJAMIN BALINSKY

United Service for New Americans and City College of New York

THE Vocational Services Department of the United Service for New Americans aids recent immigrants to achieve vocational adjustment. There is no established testing program, but outside testing facilities have been utilized on occasion. The question arose about whether or not to increase the utilization of tests for the recent immigrants. Ordinarily, the matter would have been directly answered on the basis of precedent that tests are widely accepted by counseling agencies. However, since the Vocational Services Department has recent immigrant clients, the matter of testing them more regularly was enmeshed in the broader problems of test validity and interpretation.

Design of Study

It is known that the tests employed in counseling have been standardized on American-speaking and American-accultured populations. The recent immigrants not only do not understand the American language idiom well but have had extraordinary personal experiences that make a testing program for them one that requires careful study. It was decided that the design of the study have two phases:

1. To discover the particular needs of the clients and the counselors who serve them.
2. To try out various psychological tests and techniques.

The first task has been completed. The second phase has only begun. The first phase was accomplished by means of the following:

¹This paper is adapted from a report read at the Jewish Occupational Council Eastern Regional Conference, February 18, 1949. The writer wishes to express his sincere appreciation to Mr. William Karp, Director, Vocational Services Department, for his invaluable aid in starting and working through the Psychological Services Program.

1. Conferred with supervisors and counselors on client needs and their own.
2. Observed interviews.
3. Interviewed directly, especially the more difficult clients.
4. Attended and participated in administrative staff meetings.
5. Studied case records.
6. Conferred with representatives of the Family Service Department because of the close working relationship with the Vocational Services Department.

The second phase will be accomplished by cooperating with outside testing facilities where the immigrants will be examined. The test results will be studied against interview, case history and vocational data, and the test battery modified from time to time as the results merit.

Immigrants as Special Problems

The question may be raised as to whether or not the recent immigrants present testing problems different from the usual client. From phase one of the study it was learned that:

1. The recent immigrant is generally older—42.5% were 40 years of age or older, 30% were 45 years of age or older; 22% were 50 years of age or older.
2. 82% had been in this country one year or less.
3. The largest number of applicants had been in business, salesmen, or office workers in Europe.
4. 11% were handicapped by general health or a specific physical impairment.
5. 70% were on relief and over 95% were known to the Family Service Department at some time.
6. Almost all the recent immigrants had more or less difficult social and personal adjustments to make at the same time they were making a vocational adjustment.
7. Counselors wanted help with understanding the personality of the particular immigrants. They felt that the immigrant clients were more difficult to understand than the American clients with whom they were familiar.
8. Counselors indicated that a routine interpretation of tests was of little value.

These findings put the immigrant in a special class that may well be compared to the so-called handicapped groups. Just as one does not proceed on the same basis with the handicapped as with the non-handicapped, the same cautions must be exercised with the recent immigrant. One would not rely upon oral tests

for the deaf or written tests for the blind, and one must seek tests that are more valid for the immigrants.

Testing the Immigrant

In testing the immigrant, we run into the issue of mechanical or dynamic testing and interpretation. Psychological tests have been generally accepted as part of the total process of vocational counseling. However, the use made of tests varies considerably from more or less routine mechanical interpretation of the results to the dynamic interpretation where predictions are based on all there is known about measurement principles, the particular test and the particular individual being tested. Where the individual has only a vocational problem uncomplicated by difficult social and personal adjustments and where the individual has had normal opportunities for development, a prediction based upon the specific test results will probably be valid. However, where this is not so, as in the instance of recent immigrants, then the prediction must be based on more factors than the specific test results.

Apropos of this issue I refer to the case of Hans K., as reported in the Jewish Occupation Council, Program and Information Service, Release #CM-9. Hans was an immigrant about 24 years of age. Tests had been given him twice. The first time they pointed up the need for psychiatric referral. The second time tests were given the statement was made that, "his general pattern of abilities has not changed and he has not improved much in abilities where learning power is involved, such as vocabulary." The statement continues, "on the basis of these test results it appeared that Hans could not profit from formal training. An occupation requiring either gross or precise manual dexterity, speed and some accuracy would be most suitable for him."

However, as a result of the psychiatrist's statement that Hans might react with a neurotic or psychotic breakdown "if he could not anticipate his unrealistic vocational aspirations," the results were reviewed. It was decided that typing and book-keeping might not be contra-indicated. Hans went on to make a good adjustment in office work and even successfully accomplished some part-time college work.

Here predictions of success were based upon specific test re-

sults without fully taking into consideration the language barrier and the emotional complications. Hans had rated an IQ of 137 on the Performance part of the *Wechsler-Bellevue Intelligence Scale* and an IQ of 111 on the Full Scale but only an IQ of 85 on the Verbal Scale, this test having been part of the battery given previously. An IQ of 137 on the Performance part shows a very high level of intelligence and indicated that the low Verbal IQ is most likely not permanent but very probably temporarily depressed. It should also have been known that Hans was interested in office work. Considering all the test results and what was known about Hans, the recommendation for an occupation requiring either gross or precise manual dexterity seems like clutching at straws. There seemed to be incomplete evaluation of the test results, especially in terms of Hans' particular background and personality. The need for a more dynamic interpretation was sharpened by the fact that Hans was an immigrant with emotional difficulties.

It has often been remarked that the immigrant does not have different problems, but rather more of the same that every one else has. But more of the same, a quantitative difference, makes eventually for a qualitative change. A person may be anxious upon occasion, but another may be always anxious. This quantitative difference makes for a different style of life, different kinds of adjustment. It calls for recognition by tests and by the counselor in evaluation and adjustment. Water will still be water at 99 degrees C. but at 100 degrees it will be steam and the properties will change. This qualitative change demands different handling. So it is with the recent immigrant. He may have a little more or much more anxiety, suffer more from the difference between his expectations and reality, have a greater tendency to conflict between the need to be independent and dependent. But because he has more, his problem is not only greater but different. He has less language facility, his home situation is less favorable, he has had fewer opportunities to make adjustments on his own here and to see their results. Because of this, also, he reacts differently.

Greater attention must be paid to the whole person in testing and evaluating him. We are making predictions on the basis of test results and these predictions must be based on all the evi-

dence. It is necessary to take into account: (1) theories and principles of measurement, (2) the standardization of the tests themselves and (3) the particular person being tested.

One of the major theories in measurement that is significantly related to the testing of the immigrant is that of the effect of the environment on present test abilities. There is ample evidence, both experimental and clinical, which demonstrates that an environment different from that of the group upon which the test was standardized will lead to results that require explanation. Since we are to predict the probability of adjustment to new situations we must include the possibility of accelerated growth when in a new environment, especially one that is more favorable for growth in the expected direction. Specifically, for the immigrant, this means we must be able to predict his ability level after a period of time. After a while when he becomes more accustomed to American ways, feels more secure and understands the language better, his actual test results may rise. We must be able to predict the approximate rise in the present test results. And this we cannot do unless we take into account all factors about the tests and the individual.

It is necessary to know the validity, reliability and the norm populations for each test in order to interpret the results on the tests. Validity is most important since, if a test does not measure what it is supposed to, it matters not how consistently it measures something else or from what population the norms were derived. Moreover, the validity of tests is not so high as to measure with infinitesimal error. Most intelligence tests have validity coefficients of from about .80 to .90 and aptitude test validity coefficients approximate .60 as the modal instance. This means that the error of prediction may be quite large for any one individual. This error can be reduced by studying all of the test results in terms of the individual's present state and background.

When it comes to aptitude tests where the validity based on groups is usually only about .60, the cautions in making predictions for an individual must be even greater. It may be necessary to add more tests to get at the patterns. It is important to make observations of the individual while at work on the tests. It is valuable to know about interests and ex-

periences of the individual. Only then can the predictions begin to approach significance.

Kinds of Tests

From present observations of the immigrant as a testing problem it seems that there are sufficient tests already available that are adaptable for the immigrants. It is not necessary to make new tests. Performance tests of aptitude can be administered with little difficulty. The language factor is minimized, the cultural factor is lessened and observations of method and behavior can be made to illumine the test results. We have a little experience already with some tests. Some of our clients were examined at the YMCA Vocational Service Center. We found that the usual paper-and-pencil mechanical aptitude tests did not give as valuable information in filling out the interview data as the performance type of test and those which measured abilities more specifically like the *Minnesota Paper Form-Board*. The performance tests which seemed good were the *Minnesota Spatial Relations Tests*, *Formboards A & B*, the *Finger and Tweezer Dexterity Tests*, the *Purdue Pegboard* and the *Placing and Turning Tests*. The *Wechsler-Bellevue Intelligence Scale* can be used effectively if the Verbal Part is properly evaluated.

The language factor is not as important as is the cultural. For instance, a direct translation of the Wechsler-Bellevue will still have peculiarly American items like George Washington's birthday, the height of the average American woman, some of the Picture Arrangement items and Picture Completion items. The paper-and-pencil mechanical aptitude tests have many items strangely unfamiliar, not only to immigrants, but to many of us. These kinds of tests are contra-indicated.

Clerical tests, like the *Minnesota Clerical Test*, may be administered to those immigrants who have interest in clerical work and are able to read and write English. The *Kuder Preference Record* seems preferable to the *Strong Vocational Interest Blank* for our groups.

For personality description, the projective tests, like the Rorschach, would be possible. The Rorschach has been successfully used for diagnostic purposes with immigrants. There is

some question as to its use for vocational purposes; that is, in terms of obtaining behavioral descriptions that would predict how a person would function in different work conditions. This use of the Rorschach deserves much more research. In fact the Vocational Services Department is contemplating the use of several projective techniques to get at the personality attributes and to indicate their functioning in terms of vocational goals.

The present norms on the tests can be used while immigrant norms are being developed. The immigrant norms, however, will have to be validated against the degree of success in training or at work. In this way scores on the immigrant norms can be related to the standard norms. The establishment of immigrant norms should be used as a statistic to improve the accuracy of prediction. But it cannot take the place of the holistic or clinical evaluation and interpretation of the test results. Finally, the test results need to be carefully integrated with the subsequent interviews by counselors.

AN INVESTIGATION OF THE PERSONALITY TRAITS OF ART STUDENTS¹

MARTIN SPIAGGIA

City College Vocational Advisement Unit

Introduction

MANY opinions have been voiced concerning the nature of artistic persons. Predominant among these is the belief that artists are emotionally unstable. Lombroso, a nineteenth century psychiatrist, is cited by Rank (27) to have advanced a theory on the "insanity of genius" which treated features departing from the normal as "pathological." Psychoanalysis also, as Rank shows, has tended either to identify the artist with the neurotic—particularly in Sadger's and Stekel's arguments—or to explain the artist on the basis of inferiority feelings, as in Adler's school of thought.

Whether there is any factual basis in this "abnormal" point of view, or whether it has been merely a manifestation of the universal tendency to ascribe weakness and idiosyncrasy to the highly gifted, has not yet been experimentally determined. The bulk of published psychological experimentation in this realm concerns itself with the relationships between artistic ability and such factors as intelligence (3, 33), perceptual facility (15), and creative imagination (17, 18, 19). Little work has been done, however, in studying the personality of the artist. Previous studies which appear relevant to the research at hand are described briefly below.

Data gathered on several hundred college students at the University of Minnesota (2) indicated no significant relationship between ability in art and introversion, submissiveness, or emotional instability. The *Bathurst Diagnostic Temperament Test* and the *Bernreuter Personality Inventory* were used; ability

¹ This study was submitted in partial fulfillment of the requirements for the degree of Master of Arts at New York University. For valuable help and criticism, the writer is indebted to Dr. Naomi Stewart, who sponsored the study

in art was measured by the *Meter-Scashore Art Judgment Test* and the *McAdery Art Test*, supplemented by the judgment of instructors.

In a study by Fleming of 84 girls at the Horace Mann School for Girls (6) rating scales were used for determining who the "artistic" girls were. These ratings were correlated with teachers' estimates of various personality traits. The coefficient of contingency between "talented in some field of art" and "personality," as rated by the teachers, was found to be .25, on the basis of which Fleming argues for a "definite tendency for those with artistic talent to possess what is commonly called personality." No explanation is given of what is "commonly called personality."

Prados, using the Rorschach, found the following common features among 20 professional artists (26): superior intelligence, fear of mediocrity and disregard for the routine problems of everyday life, strong drive for achievement and richness of the inner interests, and pronounced sensitiveness and emotional responsiveness to the outer world along with a lack of adaptability to it, the last mentioned feature tending to be counterbalanced by sound intellectual control.

Roe, in a study of 20 prominent American male painters (28, 29), found them to be sensitive, non-aggressive, emotionally passive, hard working, self-disciplined, and of superior intelligence. She found nothing in the personalities or intellectual powers of her subjects, as measured by the Rorschach and Thematic Apperception tests, that was *radically* different in a qualitative sense from those of other people. She found, however, that the type of social and sexual adaptation was of a markedly non-aggressive sort and hence rather more "feminine" than "masculine" according to our cultural stereotypes.

The object of the present study was to investigate differences in personality traits, as measured by the nine scales of the *Minnesota Multiphasic Personality Inventory*, between art students and non-art students matched with them on age and intelligence.

Population

Art Students—The subjects, all volunteers, were 50 male art students, age 18 or above, who had attended a recognized

art school in New York City (excluding commercial-art schools) for at least two years, and who intended to make art work their vocation.

Control Group of Non-Art Students—The control group was composed of 50 male subjects who were not art students, and were selected randomly from the general population in New York City, and in Rockland and Orange Counties of New York State. Some of the occupations included were hospital attendant, automobile mechanic, electrician, shoemaker, chauffeur, teacher, accountant, and graduate student.

TABLE 1
Comparison of Minnesota Multiphasic Personality Inventory Results Obtained on 50 Art Students and 50 Non-Art-Student Controls Matched on Age and Otis IQ

Variable	Art Students		Controls		D_M Art Student Mean— Control Mean		*D_M	t ratio
	Mean	SD	Mean	SD	Mean	SD		
Age (years last birthday)	24.64	6.16	24.62	5.47	+	.02	—	—
Otis IQ	111.56	10.49	112.68	10.82	+	.33	—	—
Hypochondriasis	52.56	10.06	51.66	9.00	+	.90	1.93	.47
Depression	56.74	10.79	53.16	5.28	+	3.58	1.66	2.15*
Hysteria	59.24	7.91	57.74	7.43	+	1.50	1.51	.99
Psychopathic Deviate	59.14	13.40	50.28	4.96	+	8.86	1.97	4.49†
Interest	70.10	11.82	55.92	5.86	+	14.18	1.86	7.62†
Paranoia	54.00	7.17	47.18	5.91	+	6.82	1.36	5.01†
Psychasthenia	53.88	10.26	50.45	6.54	+	3.43	1.67	2.05*
Schizophrenia	55.90	10.74	49.46	4.71	+	6.44	1.61	4.00†
Hypomania	61.38	10.93	53.32	5.56	+	8.06	1.72	4.69†

* Significant at the 5% level of confidence

† Significant at the 1% level of confidence.

Testing was conducted at the Psychology Laboratory of New York University between July, 1947, and July, 1948.

Each art student was matched with a control on the basis of chronological age and Otis IQ (25): within 3 points on age and 5 points on IQ. The mean age for Art Students and Controls was 24.6; the sigma on age was 6.2 for Art Students and 5.5 for Control. On Otis IQ, the mean and sigma for Art Students were 111.6 and 10.5; the Otis IQ. mean and sigma for Controls were 112.7 and 10.8.

Procedure

Raw scores for the nine scales of the *Minnesota Multiphasic Personality Inventory* (10, 11, 12, 13, 14) were obtained for

each subject. Standard score equivalents, or T-scores, were determined, full account being taken of the supplementary scores, that is, the Lie, Question, and Validity scores.

On each of the nine Multiphasic scales the difference in standard score was obtained for each art student and his non-art-student control matched for age and IQ. These T-score differences were distributed and the mean and sigma of each distribution of differences obtained.

An estimate of the standard error of the mean of each set of differences was then computed by dividing the standard deviation of each distribution of differences by the square root of $N-1$, thus allowing for the correlation in scores for the two groups introduced by the matching. A t-ratio was then computed for each variable.

Results and Discussion

Table 1 gives all pertinent data. As can be seen from this table, the art students were significantly higher than the controls in mean scores on the Depression, Psychopathic Deviate, Interest, Paranoia, Psychasthenia, Schizophrenia, and Hypomania Scales of the *Minnesota Multiphasic Personality Inventory*. These differences were significant at the one per cent level for all scales mentioned except the Psychasthenia and Depression Scales, where the differences were significant at the 5 per cent level.

If we can safely generalize from the findings of the present paper, these results suggest that the art student, as compared to the non-art student of similar age and intelligence, is more typically introverted, exhibits a greater tendency toward depression, possesses a tendency to disregard social mores or an inability to adjust to the outer world, and is more feminine in his basic interest pattern. Further, he tends toward over-productivity in thought and action, these being of unusual character, and also toward compulsive behavior.

Several factors must, however, be considered in interpreting these findings. Concerning the Interest Scale, on which the art students were found to score significantly higher, we must take heed of the caution by the authors that homosexuality must

not be assumed on the basis of a high score without confirmatory evidence, owing to the relatively low reliability of this scale. Burton (1) administered the Interest Scale to 20 rapists, 34 sexual inverts and 84 other delinquents. Although he found significant differences between inverts and rapists, and also between inverts and delinquents who were sexually normal, on retest of 34 cases the reliability coefficient was found to be only .70.

The fact that Interest scores are related to cultural factors (31) must also be taken into account in interpreting the Interest findings. Roe, for example, in the study previously mentioned (28, 29), interprets the "feminine" type of sexual adaptation of a group of male artists as reflecting the attempt on the part of our society to maintain one acceptable male stereotype.

The high mean on the Paranoia Scale would seem, however, to add weight to the significance of the high mean on the Interest Scale, in light of current psychoanalytic theory which stresses the partial failure of repression of homosexual tendencies in the psychogenesis of paranoia (4). Ferenczi (5) goes so far as to consider paranoia as distorted homosexuality. Henderson and Gillespie (13), however, describe eleven cases of paranoia in only four of which the etiology of the paranoia was in agreement with Freudian conceptions. They claim that the causation of paranoid conditions is probably not by any means uniform, but that type of personality is one of the commonest predisposing elements. The sensitive, introverted individual, such as was found common in the art-student group, is mentioned as one of the types particularly susceptible to paranoia.

The significantly high scores of the art students on the Schizophrenia and Psychasthenia Scales appear readily interpretable. It would seem likely that by virtue of his higher "cultural" level, the art student encounters difficulty in adjusting to the outer world and finds it psychologically necessary to turn inward, appearing introverted, and giving rise to the high Schizophrenia mean. The tendency toward compulsive behavior, shown by the art students on the Psychasthenia Scale, is due in part to the overlapping of items and the high corre-

lation between the Schizophrenia and the Psychasthenia Scales (.84 for normals, .75 for abnormal cases). It may also reflect a real tendency toward compulsive behavior on the part of the art student group.

The high mean score for the art students on the Hypomania Scale may appear inconsistent with the high mean for this group on the Depression Scale, and with the introverted pattern which seems to typify the group. It must be remembered, however, that the Hypomania Scale presumably measures overproductivity of *thought* as well as action. It seems plausible that the high Hypomania mean for the art students is accountable in terms of overproductivity of *thought*; that because of their introverted tendency they express these thoughts in symbolic forms rather than in action in the ordinary sense of the word.

The relatively high mean score of the art-student group on the Psychopathic Deviate Scale is contrary to expectations, in light of the introverted pattern manifested for this group, since behavior of the psychopathic deviate variety is usually associated with extroversive tendencies. The relatively high Psychopathic Deviate mean score is, however, consistent with the ordinary layman's stereotype of the artist. It must also be considered that while the Multiphasic appears adequate for giving a general over-all pattern of group behavior, it loses in validity when an attempt is made to interpret findings on any given scale taken in isolation.

Further caution is prescribed in interpreting the results discussed here. Owing to the preliminary status of some of the Multiphasic Scales, the overlapping of items among the various scales, the lack of experimental determination of reliabilities, the Multiphasic is still in an incomplete state of development.

Note must also be taken of the limitations of the present study with respect to sampling. The art students were all from professional art schools in New York City and cannot be taken to represent art students throughout the country. The number of cases, while sufficient to yield statistically significant differences for many of the comparisons made, is also very small,

in an absolute sense, the differences, therefore, while significant, are not highly reliable.

Summary

Differences in personality traits between art students and non-art students matched with them on age and intelligence, as measured by the nine scales of the *Minnesota Multiphasic Personality Inventory*, have been investigated. The findings reveal significantly higher mean scores for the art students on the Depression, Psychopathic Deviate, Interest, Paranoia, Psychasthenia, Schizophrenia and Hypomania Scales of the Multiphasic. These findings seem, on the whole, to be psychologically meaningful.

Owing to the selective character of the sample used and to the inadequacies of the Multiphasic as a tool for personality diagnosis, caution is indicated in interpreting these results.

Further study of the personality characteristics of different vocational and social groups is recommended. On the basis of the present findings, it would seem that investigations along such lines can afford material aid to the understanding of various social problems.

REFERENCES

1. Burton, A. "The Use of the Masculinity-Femininity Scale of the Minnesota Multiphasic Personality Inventory as an Aid in the Diagnosis of Sexual Inversion." *Journal of Psychology*, XXIV (1947), 161-164.
2. Carroll, H. A. "A Preliminary Report on a Study of the Relationship Between Ability in Art and Certain Personality Traits." *School and Society*, XXXVI (1932), 285-288.
3. Carroll, H. A. and Eurich, A. C. "Abstract Intelligence and Art Appreciation." *Journal of Educational Psychology*, XXXIII (1932) 214-220.
4. Fenichel, O. *The Psychoanalytic Theory of Neurosis*. New York: Norton and Company, 1945.
5. Ferenczi, S. *Contributions to Psycho-Analysis*. Boston: Gotham Press, 1916.
6. Fleming, E. G. "Personality and Artistic Talent." *Journal of Educational Sociology*, VIII (1934), 27-33.
7. Friedlander, M. I. "An Art Expert's Observations on Personality." *Character and Personality*, I (1932), 75-78.
8. Hathaway, S. R. *Supplementary Manual for the MMPI*. Part I. The K Scale and its Use. Part II. The Booklet Form of the MMPI. New York: Psychological Corporation, 1946.

9. Hathaway, S. R. and McKinley, J. C. "A Multiphasic Personality Schedule (Minnesota): I. Construction of the Schedule" *Journal of Psychology*, X (1943), 249-254.
10. Hathaway, S. R. and McKinley, J. C. "A Multiphasic Personality Schedule (Minnesota): II. A Differential Study of Hypochondriacs." *Journal of Psychology*, X (1940) 255-26.
11. Hathaway, S. R. and McKinley, J. C. "A Multiphasic Personality Schedule (Minnesota): III. The Measurement of Symptomatic Depression." *Journal of Psychology*, XIV (1942), 25-34.
12. Hathaway, S. R. and McKinley, J. C. *Manual for the Minnesota Multiphasic Personality Inventory*. New York: Psychological Corporation, 1948.
13. Henderson, D. K. and Gilligan, R. D. *A Textbook of Psychiatry*. London: Oxford University Press, 1946.
14. Hurlock, E. B. and Thompson, J. L. "Children's Drawings: An Experimental Study of Perception." *Child Development*, V (1934), 127-34.
15. Liss, F. "The Graphic Arts." *The American Journal of Orthopsychiatry*, IX (1933) 191-209.
16. Lowenfeld, V. *The Nature of Creative Activity*. (Translated by O. A. Desler.) New York: Harcourt and Brace, 1939, 272 pp.
17. Markey, F. V. "Imaginative Behavior of Young Children." *Child Development Monograph*, No. 16, 1935.
18. McClay, W. "Creative Imagination in Children and Adults." *Psychological Monograph*, LI (1949), 28-123.
19. McClay, W. and Meier, N. C. "Re-Creative Imagination." *Psychological Monograph*, LI (1950), 1-8, 116.
20. McKinley, J. C. and Hathaway, S. R. "A Multiphasic Personality Schedule (Minnesota): IV. Psychasthenia." *Journal of Applied Psychology*, V (1943) 614-624.
21. McKinley, J. C. and Hathaway, S. R. "The Minnesota Multiphasic Personality Inventory. V. Hysteria, Hypomania, and Psychopathic Deviate." *Journal of Applied Psychology*, XXVIII (1944), 151-174.
22. McKinley, J. C. and Hathaway, S. R. "The Identification and Measurement of the Psychoneuroses in Medical Practice." *Journal of the American Medical Association*, CXXII (1943), 161-167.
23. Meier, N. C. "Recent Research in the Psychology of Art." *Yearbook of the National Society for the Study of Education*, XL (1941), 379-466.
24. Meier, N. C. "Special Artistic Talents." *Psychological Bulletin*, XXV (1928), 265-271.
25. Otis, A. S. *Statistical Methods in Educational Measurements*. New York: World Book Co., 1925.
26. Prados, M. "Rorschach Studies on Artists-Painters." *Rorschach Research Exchange*, VIII (1944), 178-183.

27. Rank, O *Art and Artists*. New York: Tudor Publishing Co
1932
28. Roe, A. "Painting and Personality." *Rorschach Research Exchange*, XL (1946) 86-100
29. Roe, A. "The Personality of Artists" *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENTS*, VI (1946) 401-408
30. Schiele, B. C., Baker, A. B. and Hathaway, S. R. "The Minnesota Multiphasic Personality Inventory" *The Journal-Lancet*, LXIII (1943), 292-297
31. Terman, L. M. and Miles, C. C. *Sex and Personality. Studies in Masculinity and Femininity*. New York: McGraw-Hill, 1936.
32. Tiebout, C. and Meier, N. C. "Artistic Ability and General Intelligence" *Psychological Monograph*, XL (1936), 95-125.

THE KNOWLEDGE OF GENERAL EDUCATION OF A SAMPLE OF SYRACUSE UNIVERSITY STUDENTS AS REVEALED BY THE COOPERATIVE GENERAL CUL- TURE TEST AND THE TIME MAGAZINE CURRENT AFFAIRS TEST.

N. M. DOWNIE

The State College of Washington

M. E. TROYER and C. R. PACE

Syracuse University

DURING the academic year, 1947-1948, Syracuse University initiated an all-university self-survey, the results of which were to provide the bases for enlightened planning for the years ahead. Among the concerns of the various survey committees was an investigation of the program of general education of the University.

As a part of this study of general education, a sampling of seniors, members of the Class of 1948, and of sophomores, members of the Class of 1950, were given Form X of the *Cooperative General Culture Test* and the September, 1947, edition of the *Time Magazine Current Affairs Test*. These tests were administered late in December of 1947 and during the first school days of January, 1948. The following five Colleges of the University had students participating in the program: Applied Science, Business Administration, Fine Arts, Home Economics and Liberal Arts.

Raw test scores on the *Ohio Psychological Examination*, Form 21, were obtained for as many students as possible. Mean scores on this test were computed for each college and class. These means were tested for significant differences by means of the "t" test and the homogeneity of their variances by the "F" test. No significant difference was found between the mean score for all of the seniors and the mean score for all of the sophomores. When the mean score for each college was com-

pared with the mean of the total seniors and with that of the total sophomores, the only significant difference was found to be that the mean score of the Liberal Arts sophomores was significantly higher than that of all other sophomores.

An analysis of covariance technique was applied to the data to determine whether, if intelligence test scores were held constant, there was any difference between the total scores of the seniors and sophomores on the *Cooperative General Culture Test*. An "F" ratio of .784 was obtained. This led to the acceptance of the null hypothesis that there was no significant difference between the means of the two classes on the total scores of this test. The *Welch-Nayer Test* was used to check the assumption of homogeneity of variances of the two groups. The variances were found to be homogeneous.

When the test was analyzed by subtests and for the different classes in the five colleges, numerous significant differences appeared as shown in Table 1. The "t" test was used to test the significance of the differences between the means. Variances of each set of means were compared, using the "F" ratio. It was found in two cases where significant "t's" appeared—Current Social Problems, for the Applied Science and Home Economics students—that the real difference was caused by the variances of the two distributions.

Table 1 shows the mean total and part scores by college and class on this test. On studying this table, one sees that, in general, the students of Syracuse University achieved well above the mean on national norms for college sophomores. As a matter of fact, of the eighty-four mean scores reported in this table, only ten are below the mean on national sophomore norms.

On studying the six part-scores of the test for each college, one sees that in the College of Liberal Arts both seniors and sophomores were well above the all-university mean for each part, with the seniors and sophomores significantly different from it on Current Social Problems, History and Social Studies and Literature, and the seniors in Science. The seniors in Applied Science achieved significantly above the all-university mean in Science, Mathematics and Current Social Problems, and significantly below it in Literature and Fine Arts. The Applied Science sophomores were above the mean in Science

and Mathematics (significantly so), while on the other areas of the test, they approximated it.

The seniors in the College of Business Administration fell significantly below the mean in Literature, Science and Fine Arts and hovered around it in other areas of the test. The soph-

TABLE I
Mean Total and Part Scores by College and Class on the Cooperative General College Test

College and Class	Current Social Problems		History and Social Studies		Literature	Science	Fine Arts	Mathematics	Total
	N	Mean	N	Mean	Mean	Mean	Mean	Mean	Mean
L. A. 1948	110	146.2	149	2	143.8	133.4	45.0	24.5	1242.0
L. A. 1946	108	144.2	147	7	141.7	131.4	44.0	26.8	1228.8
A. S. 1946	40	146.5	44	5	128.7	139.4	137.0	144.2	1240.0
A. S. 1945	37	142.5	45	5	131.5	141.8	135.7	145.7	1240.4
B. A. 1946	77	144.1	43	0	129.1	125.1	137.3	23.7	1203.3
B. A. 1947	62	144.7	43	8	128.4	126.5	132.1	24.2	1199.9
F. A. 1948	51	144.9	136	1	136.9	121.9	153.8	113.7	1197.5
F. A. 1946	46	144.1	131	8	131.1	121.9	145.6	115.7	1184.2
H. E. 1948	19	149.1	116	1	124.7	27.7	39.5	113.6	1184.9
H. E. 1946	29	144.1	116	4	122.5	31.9	38.7	117.3	1190.5
Total 1948	111	142.8	45	1	125.5	20.1	43.5	23.2	217.1
Total 1946	323	141.2	42	9	121.2	30.1	138.8	126.5	212.7
National	8500	111	34	3	39.9	24.4	31.9	17.2	172.5

* Significant difference between college mean and all university mean for this class—5% level.

† Significant difference between college mean and all-university mean for this class—1% level.

‡ Significant difference between seniors and sophomores in the same college—5% level.

§ Significant difference between seniors and sophomores in the same college—1% level.

|| Based on a random sampling of approximately 8500 sophomores from those colleges whose reports were received before April 7, 1947 as reported in the 1947 National College Sophomore Testing Program, Cooperative Test Service, May 1947.

omores in the same college were significantly below the mean in the same three areas plus Mathematics and significantly above it in Current Social Problems.

Both classes in the College of Fine Arts were significantly below the all-university mean on Current Social Problems, History and Social Studies, Science and Mathematics, around the mean in Literature and significantly above it in Fine Arts.

In the College of Home Economics, both classes were significantly below the all-university mean on Current Social Problems, History and Social Studies and Mathematics, the seniors significantly below it in Literature and both classes around the mean in the other areas.

When seniors and sophomores were compared, a few inter-class differences appeared. The sophomores as a group scored significantly higher than the seniors in Mathematics and lower in Fine Arts, Literature and Current Social Problems. In the Colleges of Liberal Arts and Fine Arts, the seniors were significantly higher in Literature and Fine Arts and the Business Administration seniors in Fine Arts.

An item analysis of this test was made, using all of the papers to determine the percentage of students in each class of the five colleges who responded to each item correctly. On Part I, Current Social Problems, the students as a whole did rather well. They were best informed on items concerned with labor unions and labor activities. Other attempts to classify the items failed to show any particular area that was either very good or very poor.

Some of the Current Social Problems items, on which the students did rather poorly, are listed below. In selecting from teachers, farmers, industrial workers, white-collar workers and civil service employees, the group least likely to suffer from inflation, less than 50 per cent of the students chose the correct answer. An item which called for the knowledge that the doctrine of states' rights was used as an argument against federal antilynching legislation was known by less than 30 per cent of the students. Two other items, the period of life when the greatest incidence of tuberculosis occurs and the meaning of the term "Nisei" were likewise unknown to 50 per cent of the students. Another item which called for a definition of "nationalization of industry" was answered correctly by less than 20 per cent of the students.

A comparison of the results of the item analysis of Part II of this test, History and Social Studies, showed that on this part of the test, as on Part I, the students as a whole did quite well. As might be expected, items concerned with American history were easier than those related to European or Asiatic

history. Items related to psychology were answered very well by all of the students.

Some interesting things appeared when individual items were studied. On one item, the student selected from the following—aristocratic, autonomous, autocratic, autarchic and anarchistic—the one that best described a government controlled by the will of one man. About 50 per cent of the students answered this item correctly. The item which asked whether the election of Harding was a repudiation of the Republican Party, Ku Klux Klan, Roman Catholic Church, capitalist system or the League of Nations was likewise missed by about one-half of the students. The concept of "Balance of Power" was also unknown to about the same number of students. Perhaps one of the most elementary things missed on this part of the test was the type of government in existence in Switzerland. Twenty-five per cent of the students answered this item incorrectly.

On Part III of the test, Literature, a study of the item analysis showed that, in general, the students did poorly. Many students omitted a large number of items. There was evidence, however, that most of the items were attempted by most of the students because several items toward the end of the section were answered correctly by nearly every student.

An attempt was made to see if there were specific areas such as American literature, English literature, poetry or drama in which the students did better than in others. A comparison of the items related to American literature with those concerned with English literature showed that the students did about the same in both areas. Results on items related to Graeco-Roman literature were quite poor for all groups. When the items were studied as to whether they pertained to poetry, drama, exposition, etc., no evidence was found to show that the students did better in one of these areas than in another.

One rather interesting thing did appear from the item analysis. Included in the ninety items on literature were ten items which referred to Biblical characters or situations. Of these ten items, the students answered only two of them well. Practically everyone knew that Samson was distinguished for his strength and that the walls of Jericho came down on the blow-

ing of trumpets. Less than a quarter of the students in all of the colleges knew Lazarus as a beggar. About 35 per cent of all students were aware that St. Paul was converted to Christianity on the road to Damascus and the same percentage knew that Lot was rescued from the destruction of his city. The handwriting on the wall was recognized as occurring in the Court of Belshazzar by about 30 per cent of the students; the father and son relationship of David and Solomon was known by about 35 per cent of the students, and the fact that Joseph was sold as a slave to the Egyptians was common knowledge to only about 55 per cent of the students. On all of these items, the Liberal Arts students did only slightly better than the students in other colleges.

On the recognition of authors, the item analysis revealed that the students were very well acquainted with O Henry, Pearl Buck, Rudyard Kipling, Robert L. Stevenson, Washington Irving and Booth Tarkington; but the following authors, whose writings are more difficult and more provocative of thought, were quite unfamiliar to the majority of students—Thomas Mann, Sholem Asch, Andre Maurois, Thomas Wolfe and Aldous Huxley.

Several members of the English Department looked over this part of the test in order to judge whether each item was something that a generally educated person should know. Of the ninety items, only seven were considered as being of too technical a nature.

A study of the item analysis of Part IV, Science, showed that, except for the College of Applied Science, students were rather poorly informed about general science. Of the sixty items composing this part of the test, only three stood out as being known by almost all of the students. These items were concerned with the recognition of the metallic element found in the red coloring matter of blood, the tarnishing of silver as an example of oxidation and the major purpose of scientific investigation. One item which asked the students to select from the following the one that is not a science—organic chemistry, astronomy, bacteriology, geology and astrology—was answered correctly by as few as 60 per cent of the seniors in the College of Business Administration. Sophomores in the same college and students

in both classes of the Colleges of Fine Arts and Home Economics did only slightly better on the same item. An item concerned with the structure in animals that produces eggs was missed by almost 20 per cent of all students. The name of the gas given off by a poorly damped furnace was unknown to 50 to 25 per cent of the students in all of the colleges except Applied Science. Two items related to the use of the scientific method were rather well done by all of the students except those in the College of Fine Arts. An item on the cause of the formation of dew at night and one on the time zones of the United States were responded to correctly by about 60 per cent of all students. The law of moments was applied correctly to a problem by 50 per cent of the Liberal Arts students and by from 40 to 50 per cent of those in Business Administration, Fine Arts and Home Economics. The item which stated that osmosis is a process of 'oxidation, diffusion, absorption, reduction, or magnetic attraction' was answered correctly by about 45 per cent of the Liberal Arts students, 40 per cent of the engineers, 20 per cent of the Business Administration students, 15 per cent of the Fine Arts students and 60 per cent of those in Home Economics.

In the groups other than engineering, less than 50 per cent knew the use made of a carpenter's level. The traditional story of the bees and the birds and the flowers would have misfired with these students as less than a quarter of them knew about pollination. The general characteristics of man as a vertebrate likewise were rather obscure to these students, with less than half of them able to select from fish, sponge, oyster, lobster and insect, the animal that is most similar in structure to man.

Even the Applied Science students, who scored as a group high on this part of the test, showed that they were not generally educated in the area of science. A study of the item analysis revealed the specificity of their knowledge. In general, they did excellently on items concerned with mechanics, heat, light, sound and electricity, but, on items concerned with biology and geology, they did no better than the students in the other colleges. Several of the items concerned with the wearing of white and woolen clothing showed that the engineers transferred their knowledge of heat and light rather poorly to

actual life situations, with about 30 per cent of the students missing the items

Four members of the faculty, one each in the areas of bacteriology, chemistry, physics and zoology, were asked to look over the items on this part of the test and to consider them in the same manner as the members of the English Department were asked to treat the Literature items. The group as a whole thought that this part of the test was a rather good test of general education in the area of science. A half dozen of items were considered to be too specific to be included in a test of general education.

The item analysis of Part V of the test, Fine Arts, showed that on the whole the students were rather poorly informed in the various areas of the fine arts, except for students in the College of Fine Arts. But even in this group, unexpectedly poor results showed for many of the items.

Some of the more interesting results are noted below. Practically all of the students located the Hanging Gardens as having been in Babylon, but only from one-half to three-quarters of all the students in all colleges knew that Serge Koussevitsky was an orchestra conductor. Items concerned with contemporary fine arts were poorly answered. For example, 40 per cent or less of all students (except Fine Arts seniors, 59 per cent) recognized Thomas Benton as a contemporary American painter and less than a third of all the students identified Jacob Epstein as a modern sculptor. Salvador Dali fared a little better with from 50 to 80 per cent of the students recognizing an outstanding characteristic of his work. In music the situation was about the same. Sixty per cent or less of the students knew who wrote the *Stalingrad Symphony* and even fewer recognized the composers of *Oklahoma*.

Three members of the faculty of the College of Fine Arts went over the ninety items of this part in the same manner as faculty members treated the other areas. With a few exceptions, most of the items were thought to be concerned with things that a "generally educated" person should know in the area of Fine Arts.

The item analysis of Part VI, Mathematics, showed that this was the most difficult part of the test. The students' papers

were studied to see just how far the various groups went through the test. Students in both classes of the College of Applied Science attempted nearly all of the items. The median last item attempted was fifty-six for both classes of the College of Liberal Arts. (There are sixty items on this part of the test.) For students in the College of Business Administration, the median last item attempted was fifty-six for the seniors and fifty-nine for the sophomores. In Fine Arts this median dropped to forty-five for the seniors and to fifty-two for the sophomores, and in the College of Home Economics, the median was fifty-one for the seniors and fifty-eight for the sophomores.

A study of this item analysis showed that most students could perform the simpler arithmetical and algebraic operations. An item concerned with the extra cost involved when articles are purchased on the installment plan showed that 25 per cent of the students had no idea how to figure correctly such a common every-day problem. Forty-five per cent of the students were able to compute the annual interest on a short-term loan. A simple question involving buying and selling was solved correctly by about 25 per cent of the students outside of the College of Applied Science. Sixty-five per cent of the Applied Science students solved the problem correctly. A problem in thinking with symbols—how many minutes are there in "p" days—was also difficult, for 30 per cent or more of the students, other than engineers, could not solve it. The concept of a converse and the ability to state one was also missed by more than one-half of the students. Similarly the concept of an axiom was unknown to about 70 per cent of the students.

Course programs of a sample of thirty students, members of the Class of 1947, were analyzed to determine the number of hours the students carried in the various areas of general education. On the basis of this analysis and on the study of various course patterns as stated in the catalogs of the different colleges of the University, estimates were made of the number of credit hours the students in the different colleges carried in various areas of general education. (Estimates of the general education courses of Liberal Arts students were made entirely from the catalog.) The discussion which follows covers the five areas of general education included in the *Cooperative General Culture Test*.

The five colleges were ranked on the basis of the amount of course work in each of the several areas of general education received by the students in each college. The mean scores of the students in the five colleges on the six parts of the test were also placed in rank order. A comparison of the rank order of the number of courses taken and of the mean scores on the six parts of the General Culture Test showed that there was in general a rather close similarity between the rank order of the number of hours taken in an area of general education and the rank order of the mean score of the students in the five colleges on the part of the General Culture Test related to that area. The area which deviated most from this was the social studies. Here the Applied Science seniors, who ranked fourth in courses taken in this area, tied for first place with the Liberal Arts seniors who ranked first in the number of courses taken. In the Sophomore Class, the Business Administration students, who ranked second in the number of courses taken, tied with the Liberal Arts students, rank one in courses taken, for first place. On the History and Social Studies part of the General Culture Test, both classes of the College of Applied Science, rank four in the number of courses taken, ranked second in their mean scores on the test.

In the area of Literature, the Business Administration seniors, who were tied for lowest place in the number of hours taken in this area, were placed in third position with their mean score on the culture test. The scores of both classes of the other colleges ranked the same as the amount of literature studied with minor variations. In the area of Science there was almost a perfect relationship between the number of courses taken in science and their mean scores on this part of the test.

On the Fine Arts part of the test, both of the Liberal Arts classes, which ranked third in the number of courses taken, changed places with Home Economics, rank two. A similar switch occurred in Mathematics, in which both classes of the Liberal Arts College, rank three in courses taken, changed places with Business Administration, rank two in courses taken, on the rank of the mean score on this part of the *Cooperative General Culture Test*.

The Time Magazine Current Affairs Test.—This test was administered as an untimed test and the students were not

required to put their names on their papers. The test, as it was made up, consisted of eight parts: U. S. Affairs, Map, International, Foreign News, Canada, Science, The Arts and Personalities. However, in scoring the test, four of these parts

Map, International, Foreign News and Canada were combined into one part which was named "World Affairs." This was done chiefly because of the small number of items in each of these four parts of the test.

Mean-total and part scores for the two classes in each of the five colleges were computed. No significant differences were found between the seniors and the sophomores for the University as a whole on the total scores and sub-test scores. (The statistical techniques used here were the same as used in comparing the results of the *Cooperative General Culture Test*). When mean total scores of each college and class were compared with the all-university mean for each class, it was found that the seniors in the College of Applied Science were significantly above the mean (5 per cent level), the sophomores in the College of Business Administration in a similar situation, and that both classes of the Colleges of Fine Arts and Home Economics were significantly below the mean (1 per cent level).

When the mean scores of parts of the test were analyzed, it was found that both classes of the College of Fine Arts and Home Economics were significantly below the mean on most parts of this test. The exceptions were that both classes of Fine Arts approximated the mean on the part entitled "The Arts" and the sophomores in Home Economics were below the mean, but not significantly so in Science and The Arts. The Applied Science seniors and Business Administration sophomores were above the mean on U. S. Affairs (5 per cent level). In World Affairs, the Liberal Arts seniors were significantly above the mean (5 per cent level). In Science both classes of the engineering school were significantly above the mean at the 5 per cent level. The sophomores in the same college were significantly below the mean in The Arts (5 per cent level). In Personalities both Liberal Arts seniors and the two classes of Business Administration were significantly above the mean.

When results for the two classes in the same college were compared, the only difference between seniors and sophomores

appeared in the College of Home Economics where the seniors did significantly better on U. S. Affairs and World Affairs (both 1 per cent level) and were higher on their total scores than the sophomores (5 per cent level).

A comparison of the rank order of the mean scores of the different colleges on the various parts of this test with the number of courses taken in an area showed results similar to those obtained when scores on the *Cooperative General Culture Test* were compared with number of courses taken. The Applied Science students similarly scored high on the social studies parts of this test. The seniors ranked first on U. S. Affairs and second on World Affairs and the sophomores second on U. S. Affairs and first in World Affairs. In the number of social studies courses taken, these engineering students ranked fourth.

Summary of Findings

1. On the *Cooperative General Culture Test*, Syracuse University students ranked high according to national standards. Converting the mean scores of Table 1 into percentile scores placed the mean-total score of the seniors at the 78th percentile and of the sophomores at the 76th percentile on national norms. The average total score of students in the five colleges was well above the national average in all cases and as high as the top 11 per cent in the best case. These rather high mean percentile scores are due in part to the high scores the students made in their special areas of study and are not a reflection of a well-balanced program of general education. In the areas of the test related to the students' field of specialization, the scores averaged from the 75th to the 95th percentiles, but in the areas outside of the students' major fields the scores averaged from the 50th to the 70th percentiles.

2. When total scores on the *Cooperative General Culture Test* were compared, it was found that students in both classes of the Colleges of Liberal Arts and Applied Science scored significantly above the all-university mean. Seniors in the College of Business Administration and both classes of Fine Arts and Home Economics achieved significantly below this all-university mean.

3. Achievement in the various areas measured by this test is

definitely related to the amount and pattern of course work taken in those areas by students; and, even in the major field, students' knowledge tends to be specific to course rather than general. Students in Applied Science scored highest on the Science and Mathematics parts of the test; students in Fine Arts scored highest on the Fine Arts part of the test, etc. When the high scores on a part of the test are further examined, the specificity of the students' education is brought into sharper focus. For example, the Applied Science students did well on the items pertaining to physics and chemistry, but relatively poorly on items dealing with the biological and geological sciences and on items calling for practical applications of scientific principles to daily life.

4. In areas of study outside of their major fields, students scored relatively poorly on the test. For example, Fine Arts students scored relatively low on Science, Mathematics, and on the parts of the test related to the social sciences. Applied Science students scored relatively low on Literature and Fine Arts; Business Administration students scored relatively low on Literature, Science and Fine Arts; Home Economics students scored relatively low on Literature and Mathematics. Liberal Arts students, on the other hand, scored relatively high on all parts of the test. Similarly, the engineering students scored relatively high in the area of the social studies.

5. There is apparently no significant increment to general education during the last two years of college residence as the seniors scored no higher, or not significantly so, than the sophomores on the *Cooperative General Culture Test*.

6. On the *Time Magazine Current Affairs Test*, students in the Colleges of Liberal Arts, Applied Science and Business Administration scored relatively high, whereas students in the Colleges of Fine Arts and Home Economics scored relatively low. The total score of the seniors on this test was not significantly higher than that of the sophomores. The typical Syracuse student was able to answer about half of the items on this test correctly.

THE FULL-RANGE PICTURE VOCABULARY TEST: II. SELECTION OF ITEMS FOR FINAL SCALES¹

ROBERT B. AMMONS

University of Louisville

and

LEO D. RACHIELE

University of Denver

VOCABULARY items are among the most frequently used components of mental tests. They are, as a rule, relatively reliable and valid, take little time to administer in comparison with their usefulness, and can be given and answered in so many ways that they can often be used successfully where other items fail, as in the case of spastic children and certain aphasic adults. For general clinical use, a test should not be dependent upon the skills of reading and writing, and should avoid the ambiguities inherent in administering and scoring items calling for definitions by the testee. On the other hand, the test should be, of course, as reliable and valid as possible, should be short and easy to administer, and should have considerable intrinsic interest value.

Vocabulary items make up what are probably the best single subtests in the 1937 revision of the Stanford-Binet (11) and the *Wechsler-Bellevue Adult Intelligence Scale* (13). Terman and Merrill report an average correlation of .81 for separate age groups between the Stanford-Binet vocabulary test score and the mental age on their scale as a whole, while Wechsler states that his vocabulary subtest correlated .85 (η^2) with the total scale for the original standardization group. With these high validities in mind, a search was made by the senior author for

¹ Acknowledgment is due Professor F. Y. Billingslea, Mrs. Helen S. Ammons and Mr. Neil W. Copping of Tulane University for reading the manuscript critically and offering many helpful suggestions. The test plates and a manual with final scale norms, answer sheets and instructions for administration (1) may be obtained from R. B. Ammons.

a method of vocabulary testing which would meet the clinical criteria already outlined.

The most promising technique located seemed to be that used by Van Alstyne (12), where a child was asked to choose from among four pictures on a card the one which illustrated a particular language concept, word, or phrase. Since this test had been given only a very limited standardization and several pictures were out of date, Ammons and Huth (4) set up a new set of 16 plates and tried out a considerable number of items with a small group of children. An analysis of the results from the try-out showed that this type of test could be given quickly, was useful at least through the ages of 6 to 17, and was highly reliable and valid. On this basis, a series of studies (2, 3, 5, 6, 7) was undertaken to construct and standardize a test for all levels of verbal ability. The present paper is the first in the series reporting this work.

After a testing technique has been decided upon, at least three major problems present themselves to the constructor of a vocabulary test: (a) how to obtain items of a suitably wide range of difficulty, (b) how to select items of satisfactory representativeness of content, and (c) how to choose items valid for the estimation of differences in level of intellectual ability. It is conceivable that random sampling of all word meanings in a fairly large dictionary would provide a partial solution to these problems. Variations of this method have been used frequently. Seashore and Eckerson (10) selected a word from each left-hand page of a large dictionary, omitting prefixes, suffixes and abbreviations, and obtained a total of 1320 preliminary items. Similarly, Atwell and Wells (8) chose 100 words "by chance" from a 20,000-word dictionary. The preliminary form of the Wechsler-Bellevue vocabulary test (13) was a list of 100 words, one each chosen from the top of every fifth page in a school dictionary, omitting "obsolete, technical, or esoteric words."

In practice, random selection of vocabulary items does not work out particularly well for a number of reasons. If item selection techniques are to be employed in the choice of a final scale, randomness is lost. Word meanings should probably be used as the original population, rather than words themselves.

In a multiple-choice test, the precision of meaning tested is a function of the alternative words used with the given item. Finally, if one uses the picture vocabulary technique, certain words cannot well be represented, and item difficulty is determined to a considerable degree by the nature of the drawings themselves and the alternate drawings. For these reasons, in this test no attempt at randomness was made and our initial words were merely subjectively selected to be as representative as possible, on the basis of the pictures already available. An analysis of the results as presented later in this paper seems to justify this approach.

The problem of representativeness of content was thus handled subjectively. A suitable range of difficulty was obtained by a choice of items after testing. Several possible alternatives present themselves when one wishes to select items for validity: suitability of material can be estimated subjectively, individual item correlations with total scale score can be used, and correlations of items with outside criteria such as age or mental test results can be computed. It will be seen later that a combination of all these with several more specific criteria was actually used.

Problem

The purpose of this study was to obtain a suitable group of vocabulary items and to set up the two final forms of a picture vocabulary test, based on the 16 4-picture plates developed by Ammons and Huth (4). To accomplish this, it was necessary to find a large number of words appropriate to the cards, to try these out on a representative population, and to select those items meeting the criteria established.

Procedure

Materials.—Item selection and testing centered around 16 4-picture plates (1). With the plates already available, the next step was the discovery of a large number and variety of potentially good items to administer to the standardization group. To start with, dictionaries were checked and advanced students in psychology verbally associated with the plates as stimuli. From these sources, 243 words pertinent to the pictures were

obtained in addition to the 48 selected by Ammons and Huth (4), a total of 291. Of these, 43 were eliminated because of obvious ambiguity, probable sex differences in experience with, or regional meaning, leaving 248 items for pretesting.

Pretesting consisted of administering these 248 items with their associated plates to a small sample of children and adults varying widely in age and ability.¹ Four children were tested at each CA level 2 through 17, and four college students at each Wechsler IQ level 99-109, 112-119, 121-129, 131-138, and 140-144. For children 2 through 5 results were available from Form I. of the Stanford-Binet; those 6 through 17 were given the vocabulary test of Form I.; while full Wechsler scale results were available for the college students. These estimates of verbal ability were later used in the ranking of items by difficulty for setting up the test finally given to the standardization group. Two males and two females were tested at all but 3 of the 21 levels. The college students ranged in ability as already noted; the 2- to 5-year-olds had Binet IQ's between 90 and 110; while the school children 6 to 17 years old were judged by their teachers as being average in intelligence. Tests were administered in the same way as outlined in the procedure section for the standardization group.

After all 84 subjects had been given the appropriate intelligence test and the picture vocabulary test, the number of correct answers was tabulated for each item by age levels, and moving averages were calculated using five successive points. Per cent passing was estimated, as in Ammons and Huth's study, on the basis of these moving averages between ages. Twenty-two more items were eliminated, either because they discriminated poorly between successive age levels, or because there were too many items passed by 50 per cent of the subjects at a given level.

The resulting 226 words, including 33 remaining from Ammons and Huth's 48, were then listed by plates and by difficulty level, difficulty level being the estimated MA at the 50 per cent passing point, in terms of the intelligence tests given. The items in order of difficulty were:

¹ Thanks are due Mr. William L. Miller and Mr. Alvin Yordy, principals in the Denver Public Schools, and Mr. Gene Gullette of Englewood High School, for making subjects available.

Plate 1: pie, window, dessert, vegetable, human, seed, pane, sill, ventilation, agriculture, anti-socialness, transparent, rectangular, translucent, culinary, sector, illumination, intimidation, segment, depredation, physiognomy, egress.

Plate 2: wagon, dancing, teacher, phonograph, partners, athletes, transport, competition, revelry, terpsichorean, ebullience.

Plate 3: car, fight, boxing, counter, pump, customer, paying, clerk, fuel, sale, sport, purchase, gauge, merchant, competition, recreation, petroleum, retaliation, replenishment, pugnacity, conveyance, aggressiveness, transaction.

Plate 4: chimney, park, shrubbery, panels, dwelling, veranda, panorama, urban, domicile.

Plate 5: presents, island, surf, isolation, munificence

Plate 6: bird, horse, fly, wagon, transportation, insect, conveyance, antiquated.

Plate 7: race, catching, uniform, sport, discussion, skill, passion, affection, flight, impact, amour, dialogue, discourse.

Plate 8: house, clothes, firecracker, basket, music, laundry, clean, explosion, sudden, garment, neglect, dehydration, detonation.

Plate 9: farm, manufacturing, skyscraper, landscape, currency, industrial, pecuniary, tranquillity, agrarian.

Plate 10: chair, cup, spoon, furniture, razor, thermometer, steel, refreshment, liquid, mercury, container, grooming, beverage, centigrade, tonsorial.

Plate 11: clock, circle, numbers, locket, engraving, lobe, sentiment, appendage, chronometer, pendant.

Plate 12: food, meal, afraid, hot, fear, startling, nutrition, perspiration, tattered, vagabond, gorging, poverty, glutton, illegality, felony, humid, vagrant, coercion, mastication, destitute, gourmand, itinerant, insatiable, repast, corpulence, sudorific, mendicant.

Plate 13: telephone, accident, crying, cheerful, collision, destruction, vehicles, mishap, portrait, transmitter, sympathy, propulsion, communication, consolation, condolence, negligence, bereaved, lacrimation, deleterious.

Plate 14: policeman, safe, uniform, listening, broadcast, danger, protection, authority, disaster, gravitation, catastrophe, constabulary, fortuitous.

Plate 15: bathtub, bed, chair, newspaper, operation, illness, anaesthesia, cleanliness, aseptic, crisis, leisure, immersion, recumbent, somnolent, displacement, perusing, supine.

Plate 16: airplane, train, propellers, locomotive, intersection, harbor, aviation, altitude, marine, fuselage, nautical, roadstead.

Subjects.—The test was administered to 600 white American-born subjects ranging from age two to thirty-four years inclusive. Table I shows the number of subjects of each sex tested at each age level. Numbers are not equal because correct grade placement was a primary control, rather than age, with the

school children. Thirty were tested at each grade level but 11 18-year-olds were discarded because it did not seem possible to obtain a reasonably unbiased sample at this age level.

Between the ages of 2 and 17 inclusive, the subjects were selected by age or grade levels with respect to the fathers' occupations in direct proportion to a ten-group socio-economic breakdown as presented in the index of gainfully employed white males of the 1940 United States census (14). Where

TABLE 1
Number and Average Chronological Age of Subjects Tested at Each Chronological Age Level in the Present Standardization Group

Age level	Males		Females		Total	
	N	Mean age*	N	Mean age	N	Mean age
2	15	2.5	15	2.5	30	2.5
3	15	3.5	15	3.5	30	3.5
4	15	4.5	15	4.4	30	4.4
5	15	5.4	15	5.4	30	5.4
6	11	6.5	13	6.6	24	6.6
7	8	7.3	16	7.4	24	7.4
8	21	8.4	12	8.5	33	8.5
9	15	9.5	11	9.5	26	9.5
10	17	10.5	19	10.5	36	10.5
11	10	11.4	16	11.5	26	11.4
12	16	12.4	16	12.4	32	12.4
13	20	13.5	14	13.4	34	13.5
14	9	14.5	12	14.3	21	14.4
15	19	15.5	17	15.5	36	15.5
16	11	16.4	15	16.4	26	16.4
17	16	17.4	15	17.5	31	17.5
18-34	60	25.3	60	24.6	120	25.0
Total.	293		296		589	

* Years.

numbers per age level were below one subject, age levels were combined.

For the adult group, males and females were separately considered in direct proportion to the occupational status of white males and females between the ages of 18 and 34 as given in the census reports (14). The urban sample was obtained in the Denver area from private, parochial, and public schools; business establishments; amusement parks; and homes. The rural sample was secured from rural districts in Colorado and Nebraska. More detailed information about the sampling controls is given in articles dealing with the subgroups of the standardization population (3, 5, 6, 7).

Testing.—Preschool-age children were tested in their homes or in special rooms at day-care centers; school children were brought to rooms provided by the schools for testing; and adults were tested in their own homes, in testing rooms of industrial firms, in parks, or in a church office. All subjects were given an intelligence test and the standardization picture vocabulary test of 226 tentative items. The full Stanford-Binet, Form L, was given from ages 2 to 5, the Stanford-Binet vocabulary test from ages 6 to 17, and the Wechsler vocabulary to adults. Standard administration procedures were followed for each test (11, 13). The picture vocabulary was given first to all groups but the adults.

Since testing was done by a number of examiners, a detailed procedure including a set of instructions was set up. The subject was seated opposite the examiner, with plates and recording sheet out of sight. The session was started by asking for personal information, such as name, age, and occupation of head of family or own occupation. The subject was told he was to be asked some questions that he could answer by pointing to one of the four pictures on a plate. It was explained that some items would be too hard for him, and that he should not guess, but just say "I don't know." Doubtful items were checked by asking the subject why he made a certain choice, asking him to define the item verbally, or repeating the item later. This seemed to discourage guessing almost completely.

Items were scored right or wrong, and testing proceeded on a given plate until three successive items had been failed and three successive items passed. This was considered sufficient, since the items had been arranged in order of difficulty after pretesting, and items beyond the three-consecutive-pass and fail levels could reasonably be assumed to be passed or failed. In order to maintain rapport, the tester was free to introduce easier words at any point. Testing was started on successive plates at the subject's mental level as estimated from responses to preceding plates.

Item selection.—After the 589 subjects had been tested with the 226-word preliminary scale, an item selection was made. As a first step, all correct responses were tabulated by age, sex of subjects, and item. Words below the three-consecutive-pass level for each individual on each plate were considered as passed

and those above the three-consecutive-fail level as failed. In order to make all values of passes comparable, per cent passing was calculated from number passing and number theoretically attempting.

A CA or adult index number was found corresponding to the 50 per cent passing point for each item for the whole group, by interpolation if necessary. For example:

Word	CA levels			
	8	9	10	11
	Per cent passing			
Shrubbery	21	31	56	69

The point where 50 per cent would pass lies between CA's 9 and 10, actually at 9.8 by interpolation. CA's were used in calculating the 50 per cent passing point through age 17, while index numbers were assigned to six adult levels set up on the basis of Wechsler vocabulary scores. A word with a rating of A1.5 would have been passed by less than half of the lowest 20 adults (A1) and more than half of the next to lowest 20 adults (A2). A word with a rating of A6.5 would have been passed by less than half of the highest 20 adults (A6). Thus, relative difficulties were computed for all words in terms of 50 per cent passing points and were indicated by CA or adult index number. CA 17 and A3 were considered to be equal levels, and difficulties can be figured from below 2 to A6 in one series on this basis.

Items were rejected for the following reasons: (a) inadequate discrimination in per cent passing between successive age levels, (b) regional meaning, (c) sex difference in difficulty, (d) ambiguity of denotation, (e) same item already used with another plate, (f) too many words at a given age level.

(a) Words were thrown out where nearly the same number of subjects passed on several successive age levels, or an item was harder for a more advanced group. For example:

Word	CA levels				
	8	9	10	11	12
	Per cent passing				
Gauge	24	31	61	50	72

Words eliminated on this basis were: pane, gauge, veranda, affection, neglect, landscape, startling, tattered, vagabond, il-

legality, vagrant, destitute, transmitted, disaster, aseptic, leisure, recumbent, marine, physiognomy, conveyance, detonation, grooming, and illness. It will be noted in the following listings that several words were rejected for more than one reason.

(b) The following words were eliminated because of potentially varying difficulty depending on regional experience differences: urban, grooming, roadstead, partners, petroleum, aseptic, aviation, altitude, marine, fuselage.

(c) A separate tabulation was made of the number of males and females above and below the 50 per cent passing point for each word. Where there were marked discrepancies in item difficulty between the sexes, a chi-square test (9) was run. Apparent differences between the sexes at beyond the one per cent level were noted in the case of the following words: detonation, aviation, altitude, fuselage, partners, tattered, vagabond, illegality, vagrant, and destitute. These were eliminated, although it is realized that such marked sex differences would of course occur a number of times by chance in this large a number of words.

(d) Several words were rejected because they potentially referred to two different drawings on the same plate: customer, boxing, competition, catching, flight, glutton, illness, partners, and merchant.

(e) "Sport" and "uniform" were tried out on two different cards and the better card-word combination on the basis of the other criteria was kept.

(f) It was decided to have 10 words at each level from below 2 to 5 years, 8 at each level from 6 through 16, and 8 at each adult level 3 through 6, or a total of 170 words in the final scale. Where there were too few words, as at levels 2, 4, 5, 8, 11, 12, 14, A3, and A5, words were borrowed from adjacent levels. When a minimum number of words had been assigned to each level below—2 to 16 and A3 to A6, the surplus words were eliminated in the order that they failed to meet the other criteria. It should be noted in this connection that several of the criteria were only relative and subjectively applicable to begin with, and this final process of eliminating on the basis of an oversupply of words at a given level led to further qualitative differences between the items used with the various plates.

The final step was to divide the 170 items into two forms

equal in length and as equal in difficulty as possible. The words were therefore arranged in order of difficulty without respect to plate, and assigned in groups of four, the first and fourth of each group going to Form A, and the second and third to Form B.

Results

Following are the 85 words finally chosen for Form A, with their difficulty levels indicated:

Plate 1: pie (1.7), window (1.7), seed (6.5), sill (6.7), transparent (13.3), rectangular (14.7), sector (16.0), illumination (16.0), culinary (17.2), egress (A6.3).

Plate 2: athletes (8.6), competition (15.0), revelry (A4.0), ebullience (A6.4).

Plate 3: counter (4.0), pump (4.4), clerk (6.4), sport (7.6), recreation (10.8), pugnacity (16.9), replenishment (A3.1), retaliation (A4.1).

Plate 4: shrubbery (9.8), dwelling (11.7).

Plate 5: surf (12.5), isolation (12.9).

Plate 6: horse (1.5), wagon (2.3), insect (6.7), transportation (8.6), antiquated (A3.8).

Plate 7: discussion (7.7), skill (10.9), amour (13.8).

Plate 8: firecracker (2.7), clothes (3.0), explosion (4.9), clean (5.5), dehydration (A4.3).

Plate 9: farm (4.1), currency (12.2), tranquillity (16.5), agrarian (A6.2).

Plate 10: furniture (4.4), steel (6.0), refreshment (6.2), liquid (7.3), container (9.5), centigrade (14.5).

Plate 11: clock (1.6), locker (3.0), numbers (3.4), engraving (9.8).

Plate 12: hot (5.2), fear (7.4), nutrition (10.4), gorging (12.8), poverty (13.9), mastication (A2.6), itinerant (A4.5), coercion (A4.6), corpulence (A5.5), insatiable (A5.6).

Plate 13: telephone (2.1), crying (2.9), accident (3.0), vehicles (9.5), destruction (10.0), portrait (10.2), communication (10.6), consolation (13.4), negligence (14.3), bereaved (15.4), deleterious (A6.2).

Plate 14: danger (5.6).

Plate 15: bed (1.6), newspaper (2.5), anaesthesia (11.7), immersion (14.6), displacement (A5.0), perusing (A5.0).

Plate 16: propellers (3.7), harbor (8.1), locomotive (8.2), nautical (16.5).

The following 85 words were chosen for Form B:

Plate 1: vegetable (3.8), human (4.4), dessert (4.5), agriculture (10.7), anti-socialness (13.2), segment (15.0), intimidation (16.6), translucent (A2.5), depredation (A4.0).

Plate 2: phonograph (3.3), transport (8.4), terpsichorean (A6.0).

Plate 3: car (1.6), fight (2.8), paying (6.0), customer (6.3), fuel (7.5), sale (7.9), purchase (10.4), transaction (14.6), aggressiveness (A3.6).

- Plate 4: panels (13.9), domicile (A4.0).
 Plate 5: island (5.3), munificence (A5.7).
 Plate 6: bird (1.6), fly (2.5), conveyance (14.5).
 Plate 7: passion (12.5), impact (13.5), dialogue (13.6), discourse (A4.5).
 Plate 8: music (3.0), laundry (4.7), sudden (9.1), garment (9.8).
 Plate 9: manufacturing (7.2), skyscraper (7.8), industrial (10.0), pecuniary (A4.9).
 Plate 10: spoon (1.8), razor (3.0), thermometer (4.1), mercury (10.7), beverage (10.9), tonsorial (A4.4).
 Plate 11: circle (2.7), sentiment (13.9), lobe (15.5), chronometer (15.7), pendant (17.7).
 Plate 12: meal (3.9), perspiration (9.6), humid (14.7), felony (16.7), gourmand (A4.6), repast (A5.2), mendicant (A6.3).
 Plate 13: cheerful (6.8), collision (7.4), sympathy (9.6), mishap (11.1), propulsion (13.3), condolence (16.2), lacrimation (A6.3).
 Plate 14: policeman (2.5), listening (5.3), broadcast (5.9), uniform (6.2), safe (6.5), protection (6.7), authority (10.4), gravitation (11.8), catastrophe (12.0), constabulary (A3.2), fortuitous (A6.4).
 Plate 15: bathtub (1.6), operation (3.1), cleanliness (8.7), crisis (12.5), somnolent (16.2), supine (A5.5).
 Plate 16: train (1.5), airplane (1.8), intersection (8.5).

The point levels given with the words should be considered only as indices of difficulty, since actual average ages within age groups were not used in their calculation. The average level of Form A is 10.7 and that of Form B is 10.5. It can be seen that the forms are closely comparable in difficulty for the whole group.

Rough analyses of the incidence of parts of speech and of content areas were made for both forms combined. There are 18 words which are direct derivatives of relatively common verbs, 125 nouns, and 27 adjectives. Designating content areas arbitrarily, there are 30 words of home or domestic import, 38 referring to nature or science, 60 relating to social processes, 11 commercial, 14 personal feelings, and 17 not readily classifiable in this scheme. It would seem that the test puts a premium on the knowledge of names referring to society and social activities.

Discussion

To the extent that the occupational groups in the Denver area and a small rural area in Nebraska are typical of those in the United States as a whole, norms from this test can be considered

to be representative. There is, of course, some bias, as in all results based on controlled samples, but the sample is controlled at least as adequately as Wechsler's, if not more so. In any case it provides an excellent basis for item selection.

The words finally chosen cover the range of verbal ability thoroughly, and discriminate well between ability levels as found in different age groups. Later papers show that the two test forms made up of these words intercorrelate highly, and correlate well with other intelligence tests. The approximate age placement of items is only intended to facilitate the testing mechanically, as the test is actually a point scale. Norms for a general white population (3, 6, 7) and for certain population subgroups (2, 5) will be given for both forms in later papers.

From a practical point of view the promise of the test is well borne out. Proficient testers were able to test three or four children an hour with both the picture vocabulary test and the 1937 Stanford-Binet or the Wechsler vocabulary test. A high interest level was in evidence on the part of most of the testees. It seems from the above that it has been possible to construct a vocabulary test satisfactory for testing persons unable to speak or verbalize well.

Summary

Ammons and Huth (4) showed that it was possible to construct a picture vocabulary test of high reliability and validity. The present paper reports the procedure whereby items for such a test based on their 16 plates and covering the age levels from 2 to 34 were obtained and validated. The general procedure was as follows:

1. A set of 243 new items appropriate to the plates was listed and 48 of Ammons and Huth's final items were retained.
2. Of these 291 items 43 were eliminated by group discussion. A preliminary validation check was made on the remaining 248 words, and 226 were retained.
3. These 226 items were used to test 589 white American-born subjects ranging in age from 2 to 34 years. The sample was controlled by age levels for parents' occupation or own occupation, age-grade placement in school, and sex.
4. On the basis of this standardization testing, 56 items

were eliminated because of regional bias, failure to discriminate between successive age levels, too many items at a level, sex differences, ambiguity of picture denotation, or duplication of words on different cards.

5. The remaining 170 items were divided into two equal-length forms which were found to be almost identical in difficulty.

REFERENCES

1. Ammons, R. B. and Ammons, Helen S. *The Full-Range Picture Vocabulary Test*. New Orleans: R. B. Ammons, 1948.
2. Ammons, R. B. and Agüero, A. "The Full-Range Picture Vocabulary Test: VII. Results for a Spanish-American School-age Population." *Journal of Social Psychology*
3. Ammons, R. B. and Holmes, J. C. "The Full-Range Picture Vocabulary Test: III. Results for a Preschool-age Population." *Child Development*, XX (1949), 5-14.
4. Ammons, R. B. and Huth, R. W. "The Full-Range Picture Vocabulary Test: I. Preliminary Scale." *Journal of Psychology*, XXVIII (1949), 51-64.
5. Ammons, R. B. and Manahan, N. "The Full-Range Picture Vocabulary Test: VI Results for a Rural Population." *Journal of Educational Research*, to be printed.
6. Ammons, R. B., Arnold, P. R. and Herrmann, R. S. "The Full-Range Picture Vocabulary Test: IV. Results for a School Population." *Journal of Clinical Psychology*, VI (1950), 164-169.
7. Ammons, R. B., Larson, W. L. and Shearn, C. R. "The Full-Range Picture Vocabulary Test: V. Results for an Adult Population." *Journal of Consulting Psychology*, XIV (1950), 150-155.
8. Atwell, C. R. and Wells, F. L. "Wide Range Multiple Choice Vocabulary Tests." *Journal of Applied Psychology*, XXI (1937), 550-555.
9. Lindquist, E. F. *Statistical Analysis in Educational Research*. Boston: Houghton Mifflin Co., 1940.
10. Seashore, R. H. and Eckerson, Lois D. "The Measurement of Individual Differences in General English Vocabularies." *Journal of Educational Psychology*, XXXI (1940), 14-38.
11. Terman, L. M. and Merrill, Maude A. *Measuring Intelligence*. Boston: Houghton Mifflin Co., 1937.
12. Van Alstyne, Dorothy. *Van Alstyne Picture Vocabulary Test for Pre-school Children*. Bloomington, Ill.: Public School Publ. Co., 1929.
13. Wechsler, D. *The Measurement of Adult Intelligence*. (3rd Ed.) Baltimore: Williams and Wilkins, 1944.
14. *Sixteenth Census of the United States, 1940, Population*. U. S. Bureau of Census, U. S. Govt. Printing Office, 1940, et seq.

DOES FACE VALIDITY EXIST?

SIDNEY ADAMS

U. S. Civil Service Commission

FACE validity, in this paper, has the meaning of "appearance of validity" in the language of Mosier. Mosier (6, p. 192) says:

In this usage, the term 'face validity' implies that a test which is to be used in a practical situation should . . . appear practical, pertinent, and related to the purpose of the test . . . it should not only *be* valid, but it should also *appear* valid. This . . . is not validity in any usual sense . . . [but is] an additional attribute of the test which is highly desirable in certain situations.

This paper attempts a measurement of face validity by having a group of Federal government workers judge the extent to which seven tests possessed *true* validity. The analysis of the results attempts to answer two questions:

- (1) Does face validity exist in a form that can be reliably measured?
- (2) What relationship does face validity bear to true validity? Is a test with the appearance of validity likely to be one with actual validity?

Partial answers to both questions are to be found. Dr. Thelma Hunt had done unpublished research on the guessed validity of general psychology examinations and its relationship to true validity. Smith (8) had students evaluate seven types of examinations (essay, true-false, etc., as to their suitability for determining the grade in an education course. On the average, each student's rank of the validity of the test types correlated +.31 with any other student's estimate of validity. This indicates that under the conditions of the experiment, face validity of test-type, with the same subject-matter for all test types, is measurable, but not very reliably measurable.

¹ A discussion of the relationship between Face Validity and True Validity for the members of a group who tried out an experimental test battery.

The literature concurs in finding face validity a poor indicator of true validity. This has been pointed out by Mandell (4) and Mosier (6). O'Rourke (7) had judgments made on proposed tests for the postal service. He demonstrated that one test with great face validity possessed little or no true validity in a tryout and statistical validation which followed the judgments.

The subjects for this study were 39 members of the Personnel Department of the United States Veterans Administration. Their salary grades ranged from CAF 5 to CAF 12 (\$2634.80 to \$5905.20). It is probable that all members of the group possessed considerable knowledge of test methods.

The individuals participating in the study took eleven tests during two half-day sessions. To reduce interference with work, one session for each group of approximately 20 persons was held on one day, followed by a second session on the next day. During the second session of each of the two groups, testing was suspended. Each individual in the group was asked to rank the first seven of the tests, on the basis of their desirability for use in the selection and promotion of people for personnel jobs of the kind and level held by members of the group. Each examinee was told to write, on a sheet of paper, a code number which was used as his designation. These numbers provided anonymous identification throughout the study. The examinees were then told to write, in the time-order in which they had been taken, the names of the first seven tests of the series. These were, in order:

Administrative Judgment Test
Interpretation of Data (Graphs) Test
Vocabulary Test
English Expression Test
Contemporary History Test
Personality Estimates Test
Word Identification Test

The *Administrative Judgment Test* presents, for each question, a situation or problem in business or government organization or procedure. Five solutions are offered for each problem. The examinee is asked to choose the best of the five. The *Interpretation of Data Test* requires the examinee to read and

interpret graphs and tables which show economic and social trends. The content of the third and fourth tests, *Vocabulary* and *English Expression*, is more or less self-explanatory. The *Contemporary History Test* is a factual examination on national and world events between 1915 and 1948. It had more "background" questions and fewer straight news questions than do most current events tests. The introduction to the *Personality Estimates Test* describes the personality traits of five individuals. Each question then describes a certain action or states a certain opinion. The examinee was asked to indicate which of the five imaginary individuals would most probably have taken the action or held the opinion. The *Word Identification Test* was a type of vocabulary test in which the examinee was required to identify a particular word needed to complete a sentence. The initial letter of the word, and the number of letters in the word, were given.

The examiner described each test briefly, in order to recall all tests to the examinees.² At the time of the rating of the tests, a show of hands indicated that all examinees had reached the sixth test, *Personality Estimates*. Those who had not reached the *Word Identification Test* were asked to look at the sample questions for this test. The various tests were not separately timed, hence, at the time of the rating of the tests, the examinees had reached different tests in the battery.

The examinees were asked to consider which one of the seven tests was the best for selection for, or promotion to, personnel jobs in the Veterans Administration, or similar personnel jobs. The tests were to be ranked according to their present state; no allowance was to be made for possible improvements in the tests. The best test in each list was marked "1 best", the next best as "2", and so on to "7" for the poorest test. Tied ratings were to be reconsidered; the examinee was to break the tie arbitrarily if tests appeared tied after reconsideration.³ Examinees were cautioned that "face validity"

² The first seven, rather than all eleven, tests were used. This was done to allow the tests to be rated at a convenient time in the schedule. Also, there would be probably considerable confusion among the judges in comparing, by recall, as many as eleven tests.

³ The examiner assigned a "4" or average rating to one omitted test. Two tests remained tied, presumably after reconsideration by the rater. The examiner broke the tie by tossing a coin.

was not the major consideration in the selection of a good test; that a test might sometimes be a poor selective or promotional aid in spite of apparent validity.

The distribution of the ranks assigned each test is shown in

TABLE 1
Frequency Distribution, Mean and Variability of the Rank in Face Validity of Seven Tests by Veterans Administration Personnel Workers

	1	2	Rank 3	4	5	6	7	M	σ
Administrative Judgment		19	7	6	4	1	1	2.18	1.52
Interpretation of Data		2	14	13	2	0	5	3.28	1.71
Vocabulary		1	1	4	8	12	9	4.85	1.37
English Expression		4	1	7	14	4	7	4.08	1.58
Contemporary History		3	3	4	3	7	9	4.92	1.92
Personality Estimates		10	10	2	0	4	1	3.74	2.51
Word Identification		0	3	3	8	11	7	4.95	1.45

TABLE 2
Horst's Reliability for Rated Face Validity of Tests

	a	b	c	d	e	f	g	h
	n	ΣX	ΣX^2	$\frac{\Sigma X}{n}$	$\frac{\Sigma X^2}{n}$	$\left(\frac{\Sigma X}{n}\right)^2$	c-f	$\frac{g}{(n-1)}$
Admin.	39	85	275	2.18	7.05	4.75	2.30	.0605
Interp.	39	128	534	3.28	13.69	10.76	2.93	.0771
Vocab.	39	189	989	4.85	25.36	23.52	1.84	.0484
English	39	159	745	4.08	19.10	16.65	2.45	.0645
Contemp.	39	192	1088	4.92	27.90	24.21	3.69	.0971
Personal.	39	146	792	3.74	20.31	13.99	6.32	.1663
Word Id.	39	193	1037	4.95	26.59	24.50	2.09	.0550
Sum.		1092	5460	28.00	140.00	118.38	21.62	.5689
				A		B		C

Table 1. The mean and standard deviation of the rank assigned each test is also shown in this table.

In terms of Horst's (2) formula for reliability, the reliability of face validity ratings amounts to .911. This is shown in Table 2, which is arranged according to Horst's work sheet for his formula. Thus, it appears that the measure of face validity used does have reliability.

(n = no. of ratings per test) (X = raw ratings)
 (N = no. of tests)

$$r = 1 - \frac{C}{B - \frac{A^2}{N}} \quad r = 1 - \frac{.5689}{118.38 - \frac{784}{7}} = .911$$

TABLE 3
Variance Analysis of Ratings of Tests

(1) Test	(2) Variance of rank of test	(3) Deviation of the rank of test from 4, the mean rating (rank) of all tests	(4) Square of Column (3)
Administrative Judgment	2.2993	-1.82	3.312
Interpretation of Data	2.9204	-.72	0.518
Vocabulary	1.8740	+.85	0.722
English Expression	2.418	+.08	0.006
Contemporary Affairs	3.6609	+.92	0.846
Personality Estimates	6.2935	-.26	0.068
Word Identification	2.0994	+.95	0.092
	11.6288	0.00	6.374
	$\times 39$		$\times 39$
Within - Variance	843.5232	Between Variance	248.586
+ 266	3.171	+ 6	41.431
Natural log of quotient		Natural log of quotient =	
		$\frac{3.72469}{3.72469 - 1.15426} = 2.57043$	
	1.15426	$\div 2 = 1.28522, (z)$	

TABLE 4
Rank Correlations between Rated Estimates of Test Validity for Random Pairs of Examinees
 —Raters

Pair	r	sp	Pair	r	sp	Pair	r	sp
1	+.250	0.356	7	-.107	0.374	14	+.036	0.377
2	-.286	0.349	8	-.179	0.367	15	-.643	0.230
3	+.393	0.324	9	+.750	0.174	16	+.286	0.349
4	-.071	0.376	10	+.714	0.194	17	+.071	0.376
5	+.321	0.342	11	-.214	0.362	18	-.071	0.376
6	+.608	0.246	12	+.607	0.247	19	+.393	0.324
			13	+.518	0.284			

This was confirmed by a variance analysis following the method of Mills (5) which showed the between-variance greater than the within-variance at a probability within the one per cent level. z was equal to 1.28. These calculations are shown in Table 3. It thus appears that face validity is a definite entity, whether or not face validity is related to true validity.

Some tests do appear to this group of examinees to be more valid than others.

A different approach to this problem is by determining whether face validity shows any reliability. This is demonstrated by showing the rank correlations between raters. Of the 39 raters, 38 were paired in random pairs. The final digits of logarithms in corresponding positions on successive pages in a logarithmic table were used to determine the pairing of the individuals, each of whom had a code number. Rank correlations (ρ) were determined for each pair. These correlations are shown in Table 4. For computation of the standard error of ρ , see (1, p. 123).

TABLE 5

	S	S ²
Administrative Judgment	85	7225
Interpretation of Data	128	16384
Vocabulary	189	35721
English Expression	159	25281
Contemporary History	192	36864
Personality Estimates	146	21316
Word Identification	193	37249
		180040

The mean ρ amounts to +.178, and the median to +.250. Thus the relationship between the ranks of the tests by any two individuals, while very small, is positive. A further measure of the interrelationship of the pairs can be obtained by the use of the average intercorrelation formula. The use of this formula gives an answer to this question: Do the rank correlations computed in Table 4 appear representative of all the possible correlations which could be computed between the ranks of tests for different pairing combinations of subjects? A total of 741 correlations would be possible. The mean of these correlations has been computed by the average intercorrelation formula, as used by Smith (8), and explained by Kelley (3). The formula used was—

$$r_{11} = 1 - \frac{a(4N + 2)}{(a - 1)(N - 1)} + \frac{12\Sigma S^2}{a(a - 1)N(N^2 - 1)}$$

In this study, a is 39, the number of subjects. N is 7, the number of ranks, which is equal to the number of tests. S is the sum for each test, of the square of each rank, times the number of times the test was assigned to that rank. The computation of ΣS^2 is shown in Table 5. The value of r_{11} is $+.33$. This agrees fairly well with the observed values of the 19 rank correlations. Thus, by the use of both variance analysis and correlation, the measurable existence of face validity has been shown.

What is the relationship of face validity to actual validity? The true validity of the tests was measured by correlating the

TABLE 6
Relationship of True Validity to Face Validity

(1) Test	(2) Face Validity	(3) True Validity (Grade) (Criterion)	Rank Col (2)	Rank Col (3)	(4) True validity (others averaged judgment criterion)	Rank Col. (4)
Administrative Judgment.	2.18	.52	1	1	.50	2
Interpretation of Data .	3.28	.16	2	5	.31	4½
Vocabulary	4.85	.26	5	3	.52	1
English Expression	4.08	.13	4	6	.18	7
Contemporary History	4.02	.08	6	7	.25	6
Personality Estimates	3.74	.32	3	2	.42	3
Word Identification	4.95	.22	7	4	.31	4½

$r = +.50$ —grade criterion.

$r = +.31$ —judgment criterion.

test scores with the average rating of each participant. The ratings of each participant were made by other participants, who claimed knowledge of his work. Another measure of true validity used as a criterion was the civil service grade of the participants.

In Table 6 the face validity and both kinds of true validity are shown for each of the seven tests. The rank of the test in each of these characteristics is shown. The rank correlation of the face validity is $+.31$ with true validity determined with a judgment criterion. It is $+.50$ with true validity computed against a salary-grade criterion.

In Table 7 it is seen that the relationship of the true validity of a test to its validity estimated by one individual is very

TABLE 7
Rank Correlations between Face Validity and True Validity for Each Participant in the Test Tryout

Participant	Relationship to True Validity as Measured by			Participant	Relationship to True Validity as Measured by			Participant	Relationship to True Validity as Measured by		
	Job Performance	Grade	Participant		Job Performance	Grade	Participant		Job Performance	Grade	Participant
1	+ .062	.376	+ .286	14	-.366	.330	- .428	27	+ .295	.348	+ .607
2	- .473	.296	- .321	15	+ .295	.348	+ .643	28	-.027	.378	+ .286
3	+ .062	.376	+ .429	16	+ .295	.348	+ .536	29	+ .580	.259	+ .786
4	+ .009	.378	+ .357	17	+ .420	.314	+ .750	30	+ .420	.314	+ .536
5	+ .152	.369	+ .429	18	-.295	.348	- .035	31	-.098	.374	- .214
6	- .402	.317	- .214	19	+ .027	.378	+ .500	32	+ .420	.314	+ .536
7	+ .578	.250	+ .393	20	+ .295	.348	+ .679	33	+ .455	.306	+ .786
8	- .116	.373	- .179	21	-.027	.378	+ .393	34	+ .241	.356	+ .071
9	+ .134	.371	+ .321	22	+ .420	.314	+ .286	35	+ .384	.325	+ .750
10	- .098	.374	- .071	23	-.384	.325	-.429	36	+ .437	.309	+ .536
11	+ .170	.367	- .036	24	+ .134	.371	+ .179	37	- .063	.376	+ .036
12	+ .098	.374	+ .143	25	- .134	.371	-.071	38	+ .634	.235	+ .107
13	+ .580	.259	+ .714	26	+ .348	.335	+ .286	39	+ .187	.365	- .143

small and undependable. The numbers used in Table 7 are not the code numbers used in the examination.

Conclusions

1. Face validity appears to exist, at least for the tests, subjects and conditions described in this paper. Examinees, exposed to several tests, agreed with measureable consistency that some of the tests appeared more valid than others.

2. Wide differences often exist between the judgments made by different individuals as to which tests possess face validity.

REFERENCES

1. Dunlap, Jack W. and Kurtz, Albert K. *Handbook of Statistical Nomographs, Tables and Formulas*. Yonkers-on-the Hudson: World Book Company, 1932.
2. Horst, Paul. "A Generalized Expression for the Reliability of Measures." *Psychometrika*, (1949), 14, 21-32.
3. Kelley, Truman L. *Statistical Methods*. New York: Macmillan Company, 1924.
4. Mandell, Milton M. "Facts and Fallacies of Personnel Testing." *Personnel*, XXIV (1947), 112-115.
5. Mills, Frederick C. *Statistical Methods*. New York: Henry Holt and Co., 1938.
6. Mosier, Charles I. "A Critical Examination of the Concept of Face Validity." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, VII (1947), 191-206.
7. O'Rourke, L. J. "Saving Dollars and Energy by Personnel Research and Investigation in the Interest of the Postal Service." *Journal of Personnel Research*, IV (1926), 351-364, 433-450.
8. Smith, Francis T. "The Relationship Between Objectivity and Validity in the Arrangement of Items in Rank Order." *Journal of Applied Psychology*, XX (1936), 154-160.

ADMINISTRATION OF THE PURDUE PEGBOARD TEST TO BLIND INDIVIDUALS

JAMES W. CURTIS

Illinois Division of Vocational Rehabilitation

APTITUDE testing of the blind involves certain difficulties not always encountered in connection with normal individuals or individuals with other types of physical handicaps. Perhaps the two principal difficulties are the limitations in potential vocational placements, and the limitations in available testing instruments. Although the past decade has witnessed increasing attention to the development of suitable instruments, a substantial number of aptitude factors still present relatively difficult problems of determination, as applied to blind persons.

In successful rehabilitation and job placement, the necessity for careful evaluation increases in direct proportion to the severity of the handicap. Improvisation often becomes a necessary part of the repertoire of the psychological tester, particularly in those instances in which blindness is the handicap.

It was noted by the author, on numerous occasions, that job placement of blind individuals by the Illinois Division of Vocational Rehabilitation involved an element of finger-hand dexterity not satisfactorily measured by commonly used adaptations of standard manipulative and dexterity tests such as the *Pennsylvania Bi-Manual Work Sample* and the *Minnesota Rate of Manipulation Test*. After some trial-and-error investigation, it was determined that the *Purdue Pegboard Test* could be used, with very little special adjustment, in a quite satisfactory manner with blind individuals. It was found, moreover, that the results so obtained provided a significant addition to the results obtained from other manipulation and dexterity tests, in standard use with the blind, such as the two mentioned above.

The utility of any "standard test," in conditions of specialized use, is in inverse proportion to the complexity of the special

adjustments necessary for such use. At the same time, the fewer the necessary adjustments, the greater will be the adherence to the original standardized conditions and, consequently, the more significant will be the results from the standpoint of the original test purpose. Fortunately, the *Purdue*

TABLE 1
Purdue Pegboard Test
Norms for the Blind
(*N*=70)

Percentile	Insertion	Assembly
99	40	38
95	39	36
90	38	34
80	34	32
70	31	30
60	29	28
50	26	26
40	25	25
30	23	23
20	21	21
10	17	18
5	14	14
1	4	2

Pegboard Test may be administered to the blind with only the following deviations from standard instructions:

- a. As the examiner introduces the test, he assists the subject in manually examining the board, locating the cups, examining the pins, sleeves and washers, and identifying the rows of holes.
- b. At the start of each sequence, the tester places one pin (or one assembly) in the first hole of the row of holes to be used. In the two-hand sequence the operator places a pin in the first hole of both rows. The pin or pins so placed do not count in scoring but serve as orientation points for the blind subject. No additional deviation from the original instructions is necessary. It is desirable, however, to have the subject re-examine the sleeves and washers before proceeding with the assembly section.

Up to the present time, 70 blind subjects have been tested by the *Purdue Pegboard Test* in conformity with the instructions outlined in the above paragraph. The age range was 18 to 44 years, with the distribution of ages approximating a bell curve. There were 45 male subjects and 25 females. The IQ range was 89 to 130, with the average, 107. Each of the 70

subjects had voluntarily contacted the Division of Vocational Rehabilitation for rehabilitation. The 70 were tested in turn, according to their date of application for services. Other than this, no selective factors were in operation. Norms obtained from these 70 cases are presented in Table 1, in terms of percentiles.

The scores included in Table 1, under the designation "insertion," represent the total number of pins inserted by right hand, by left hand, and by both hands, for one trial. The scores designated as "assembly" represent the total number of pieces assembled in one trial, on the section of the test designated "assembly" by the publisher.

A study of the insertion scores in Table 1, using the 1948 Purdue Pegboard Profile Sheet for comparison, will show that the 99th percentile (blind norms) is equivalent to the 29th percentile, for industrial applicants. The 50th percentile (blind norms) is below the first percentile level, for industrial applicants. A comparison of the assembly section norms of Table 1 with the 1948 Profile Sheet, shows that the 99th percentile (blind norms) is equivalent to the 80th percentile, and that the 50th percentile (blind norms) is equivalent to the 15th percentile.

Although an insufficient period of time has elapsed to permit a statistically reliable validation of the norms for the blind, on the basis of achievement in training or employment involving finger-hand dexterity, preliminary results have indicated the strong advisability of utilizing such data as a part of the vocational testing complex.

Summary

The *Purdue Pegboard Test* was administered to 70 blind individuals, subject only to minor modifications in administrative technique. Tentative norms, based on these administrations, were determined in terms of percentiles. Incomplete results suggest a significant level of utility for measurements obtained by this technique, in vocational guidance and placement of blind individuals.

EVALUATING PSYCHOMETRIC PROFICIENCY

FRANK M. DE MAS

American Council on Education

Introduction

INDIVIDUALS who have the responsibility of training applied psychologists are often faced with the problem of evaluating the ability of their protégés to administer individual tests. There are two considerations involved. First, the evaluation of the student as compared to other students. Second, the evaluation of the student as compared to a professional standard of competency. Because of the guild-type training received, the evaluation may be highly subjective. It would seem, therefore, that an objective method of evaluating psychometric proficiency would serve as a useful supplement to the generalized subjective evaluation of the supervising clinician.

Time is important to the busy clinician. The rationalization of the two procedures that follow was made with this constantly in mind. The problem may be stated thus: Can an objective procedure be worked out which the supervising clinician can apply *routinely* in appraising psychometric proficiency? Of the two tests that follow, the first can be made in a minute or so and the second should seldom require more than three or four minutes.

Analysis of the Standard Error of Measurement

The square of the standard error of measurement, σ_1^2 , may be regarded as the variable¹ error variance of a test score. This variance has two components: the variance due to the psychometrician, σ_p^2 , and the variance not due to the psychometrician, σ_{np}^2 , as

$$\sigma_1^2 = \sigma_p^2 + \sigma_{np}^2. \quad (1)$$

¹ Errors may be classified as either variable or systematic. The present author would like to point out that this paper does not evaluate systematic error. The present method, therefore, is applicable only when an evaluation of variable error is desired. The method suggested in this paper is meaningless when only systematic error is present.

The variance σ_p^2 may be regarded as composed of two components also: the variance due to the psychometrician himself, σ_{ph}^2 , and the variance due to the situation in which the test is administered, σ_s^2 . But since the trained psychometrician is responsible for giving the test under specified conditions, we may write

$$\sigma_p^2 = \sigma_{ph}^2 + \sigma_s^2. \quad (2)$$

The variance σ_{np}^2 may be regarded as having two components: the variance due to the testee, σ_t^2 , and the variance due to the test instrument, σ_i^2 . That is,

$$\sigma_{np}^2 = \sigma_t^2 + \sigma_i^2. \quad (3)$$

It is obvious, however, that σ_i^2 is usually infinitesimal when the same test is used. When a parallel form is used interchangeably σ_i^2 may increase but usually only slightly. Since $\sigma_i^2 \rightarrow 0$ we may disregard this quantity and write

$$\sigma_{np}^2 \approx \sigma_t^2. \quad (4)$$

It follows that

$$\sigma_{I\infty}^2 = \sigma_p^2 + \sigma_t^2. \quad (5)$$

It is obvious that an unskilled psychometrician should be less reliable than a skilled psychometrician, i.e., the square of the standard error of measurement derived from test scores obtained by an unskilled psychometrician, σ_{xxu}^2 , should be larger than the square of the standard error of measurement derived from test scores obtained by a skilled psychometrician, σ_{xxs}^2 , as

$$\sigma_{xxu}^2 > \sigma_{xxs}^2. \quad (6)$$

Since (6) would be true even if both the skilled and unskilled psychometrician used the same testees, and the same test in the same situation, it follows that (6) is due to the fact that the variance due to an unskilled psychometrician, σ_{up}^2 , should be larger than the variance due to a skilled psychometrician, σ_{sp}^2 .

We should expect, therefore, $\sigma_{xxu}^2 > \sigma_{xxs}^2$ because

$$\sigma_{up}^2 > \sigma_{sp}^2. \quad (7)$$

Now, the psychometricians who standardized a particular test may be regarded as expert or skilled psychometricians.

Therefore, we may substitute the square of the standard error of measurement as published in the standardization data, σ_{xx}^2 , for the variance σ_{xx}^2 in relation (6) as

$$\sigma_{xx}^2 > \sigma_{xx}^2. \quad (8)$$

The quantity σ_{xx} , which is the published standard error of measurement for a particular test, will be used in the procedures that follow as the standard error of measurement desired from a skilled psychometrician. The square of the quantity, σ_{xx}^2 , will be regarded as the variance of a population of measures obtained by a skilled psychometrician on a single individual. It follows that the degrees of freedom for σ_{xx}^2 will be ∞ .

Criterion of Psychometric Proficiency I

Procedure:

- a) Regard the first test score, S_1 , obtained from a testee by a psychometrician as the mean of a population of such measures.
- b) Regard the second test score, S_2 , obtained by the psychometrician from the testee as a deviation from the mean.
- c) Regard σ_{xx} as the standard deviation of a normally distributed population of such measures.
- d) Criterion of psychometric proficiency, I , is attained when the second test score does not deviate significantly from the first test score, i.e., when the null hypothesis is *acceptable*.
- e) Test the null hypothesis by applying the following formula

$$x = \frac{S_1 - S_2}{\sigma_{xx}} \quad (9)$$

where x = deviation from the mean in terms of sigma as the unit.

- f) Enter the normal probability table with x and obtain the probability area, A , lying between this deviate value and the mean. Multiply this area by 2 and subtract this product from 1. If the decimal place in the remainder be moved two places to the right we then have the level of confidence at which the null hypothesis may be rejected. These operations may be summarized as follows:

$$I. C. = 100(1 - 2A). \quad (10)$$

Evaluation: The assumption given in (a) above is implicit in all test scores obtained in the clinic. It is the rule rather than the exception that only one test score of a kind is obtained from

the testee and this score is considered as an estimate of the mean of a population of such measures

The level of confidence at which the null hypothesis may be rejected is set by the supervising clinician. The severity of the criterion may be increased as the training progresses by merely setting the level of confidence for rejecting the null hypothesis at a lower point—say, from the 10 per cent L. C. to the 40 per cent L. C.

Application: Let us assume that a group of psychometricians are to be evaluated. Table 1 demonstrates the actual computation necessary. Explanation of Table 1 follows:

Col. 1: Names of evaluated psychometricians

Col. 2: The two test scores obtained by each psychometrician, (S_1 , S_2). Let these be Wechsler-Bellevue IQ's

TABLE 1
Evaluation by Criterion I

Col. 1	Col. 2	Col. 3	Col. 4	Col. 5	Col. 6
Smith	132 126	6	5.674	1.06	29%
Jones	84 86	2	5.674	.35	73%
Brown	92 108	16	5.674	2.82	1%

Col. 3: The difference between the two test scores obtained by each psychometrician.

Col. 4: The published standard error of measurement for the particular test being used; in this example the Wechsler-Bellevue test of intelligence (1). This is σ_{xx} .

Col. 5: x as defined in formula (9).

Col. 6: Approximate level of confidence at which the null hypothesis can be rejected.

The psychometricians may be compared as follows (see Col. 6): Jones is best, Smith is next best, and Brown is the poorest in psychometric proficiency in regard to the Wechsler-Bellevue test of intelligence.

If the criterion of proficiency were set at the 20 per cent L. C., then Jones and Smith passed the criterion and Brown failed the criterion.

*Criterion of Psychometric Proficiency II**Procedure.*---

- a) Regard two or more scores obtained by a psychometrician from a single testee as a random sample from a normal population of such measures.
- b) Regard σ_{xxp}^2 as an estimate of the variance of this population. $\sigma_{xxp}^2 = \Sigma d^2/n - 1$, where Σd^2 is the sum of the squared deviations from the mean of the sample in (a) and "n" is the number of scores in the sample.
- c. Regard σ_{xx}^2 as an estimate of the variance of a normal population of a set of measures obtained from the testee by a skilled psychometrician.
- d. Criterion of Psychometric proficiency, II is attained when σ_{xxp}^2 is not significantly greater than σ_{xx}^2 .

TABLE 2
Evaluation by Criterion II

Col. 1	Col. 2	Col. 3	Col. 4	Col. 5	Col. 6
Joe	133 120 128	32.19 d.f. = ∞	43.00 d.f. = 2	1.34	>5%
Tom	90 86 94 98	32.19 d.f. = ∞	26.67 d.f. = 3	1.21	>5%
Bill	92 112 96	32.19 d.f. = ∞	112.00 d.f. = 2	3.48	3%

- e. Test the null hypothesis by first applying the formula

$$F = \frac{\sigma_{xxp}^2}{\sigma_{xx}^2}, \quad (11)$$

where the d.f.² of σ_{xx}^2 may be taken at ∞ and the d.f. of σ_{xxp}^2 is $n - 1$.

Evaluation: The supervising clinician should first inspect σ_{xxp}^2 and if $\sigma_{xxp}^2 < \sigma_{xx}^2$, the psychometrician being evaluated is less variable than the skilled psychometrician—at least on the basis of this estimate of his variance—and the F test need not be made. However, if the supervising clinician wishes to know whether or not the psychometrician being evaluated is signifi-

² The degrees of freedom of σ_{xx}^2 is exactly the size of the standardization sample minus one. Since the standardization sample is usually several hundred, the error introduced by setting the d.f. always at ∞ is very, very small. The utility is that only 1 line of the F table need be used.

cantly *less* variable than the skilled standardization psychometrician he may make the F test

$$F = \frac{\sigma_{xx}^2}{\sigma_{xpx}^2}, \quad (12)$$

where the degrees of freedom are the same as in (9).

From Formula (6) it follows that we should expect $\sigma_{xpx}^2 > \sigma_{xx}^2$. The application of formula (11), applied only when $\sigma_{xpx}^2 > \sigma_{xx}^2$, will indicate whether or not the intraining psychometrician is significantly *more* variable, and therefore less reliable, than the skilled standardization psychometrician.

Application: Let us assume that a group of psychometricians are being evaluated. Table 2 represents the actual computation necessary. Explanation of Table 2 follows:

Col. 1. Names of evaluated psychometricians.

Col. 2: Wechsler-Bellevue IQ's obtained by each psychometrician.

Col. 3: The square of the standard error of measurement as published for the Wechsler-Bellevue test of intelligence, i.e., $(5.674)^2$. This is σ_{xx}^2 .
d.f. = degrees of freedom.

Col. 4: Estimated variance for a population of such measures as sampled in Col. 2. This is σ_{xpx}^2 .
d.f. = degrees of freedom.

Col. 5: Fisher's F ratio.

Col. 6: Level of confidence for rejecting the null hypothesis.

The psychometricians may be compared as follows: Tom is best, Joe is next best and Bill is the poorest psychometrician in regard to the Wechsler-Bellevue test of intelligence. From $26.67 < 32.19$, we know that Tom is less variable and, therefore, probably more reliable than even the standardization psychometrician. However, Tom is not significantly more reliable.

If the criterion of proficiency had been set at the 5% L. C., then Tom and Joe passed the criterion and Bill failed the criterion.

REFERENCE

1. Wechsler, D. *Measurement of Adult Intelligence*. Baltimore: Williams & Wilkins, 1945.

INTEREST AND PERSONALITY MEASURES OF VETERAN AND NON-VETERAN UNIVERSITY FRESHMAN MEN

KATHERINE K. FASSETT

University of Wisconsin

Fifty veterans and fifty-six non-veterans, all freshman men coming to the University of Wisconsin Student Counseling Center in 1946-48, have been investigated with respect to their interest scores on the *Strong Vocational Interest Blank* and their personality scores on the *Minnesota Multiphasic Personality Inventory*. Both the Strong and the Multiphasic are routinely administered to all students coming to the Counseling Center; Multiphasic scores are K-corrected (3), and the Strong scored on thirty-four occupations in eleven groups. The ages of the non-veterans ranged from 17 to 19 years with the median at 18; of the veterans, from 20 to 30, with the median at 22. The length of service of the veterans ranged from 24 to 72 months, the median being at 33 and a half months. All had some service outside of the continental United States. The academic classifications of Letters and Science, Engineering, and Agriculture are represented in both groups.

Interests, as measured by patterning on the Strong (1), show no significant differences between the two groups of men. Judged by the total number of A and B+ scores, the veterans have more fully crystallized interests than do the non-veterans. This difference is significant beyond the one per cent level of confidence, the veterans giving more of the high scores than do the non-veterans. Such increase in crystallization of interest with added age has been found in previous investigations (4). In the case of the groups compared in the present study, there is no overlap in age; the difference here found might consequently be expected in terms of age alone. However, the studies on which such a difference has been demonstrated have not had the factor of war experience affecting the older group, and it has

sometimes been thought that the service experience of veterans may have hindered the maturing of their vocational interests which would otherwise have come about with added age. The comparison of high scores in the present study indicates that such maturing has gone on in the veteran group, although to what extent, as compared with men of similar age in non-war years, cannot be judged from this evidence.

Multiphasic measures show no significant differences between the two groups in central tendencies on any scale; both groups have mean profiles which run close to the 50 t-score mean of the general population. The mean t-score for no scale, for either group of men, was higher than 59, or lower than 49. Greater variability for the veterans was shown on several of the scales

TABLE 1
*Standard Deviations of Multiphasic T-Scores**
Comparison between Veterans and Non-Veterans

Scale	Veterans N = 50	Non-veterans N = 56	C.R.
	S.D.	S.D.	
Hs	9.22	6.93	2.01
Pt	15.22	10.88	2.35
Mf	12.47	8.71	2.47
D	15.12	10.66	2.42

* Scales which are not listed show differences significant at $>.05$ level of confidence.

(Table 1); and the veterans appear somewhat more often than do the non-veterans in the score ranges indicative of possible personality deviations. Counting the total number of scores on all scales for each group, and computing the percentage of such total which falls at or above 75 t-score, the veterans show a greater percentage of high scores than do the non-veterans. On the Mf scale, a larger percentage of the veterans than of the non-veterans score at and above a t-score of 70. Both of these differences are significant beyond the one per cent confidence level. As the Mf scale is usually interpreted, the fact of the veterans scoring higher on the scale would indicate the presence of more feminine tendencies on their part than on the part of the non-veterans. The assumption is often held that young men of college age show some aggression to over-protection by their

mothers, and take on some feminine attributes in order to compete with the mothers. The means of both groups of men in this study run somewhat higher than the mean of the general population, but the fact that the veterans' mean is significantly above the non-veterans', can probably not be accounted for by mother relationships, since, on this basis alone, the non-veterans would be expected to run higher, inasmuch as they have recently been closer to their homes. The presence of more feminine tendencies on the part of the veterans might be due to the fact that, having been separated from considerable feminine contact for some time, they react to such contacts when entering the college situation in a coeducational institution—possibly showing an aggression towards, or competition with, the female student body which had been predominant on the campus before the return of the veterans. The fact that these young men have chosen to undertake an education rather than to get some gainful employment immediately might be the result of one or both of two tendencies commonly considered to be feminine. an interest in cultural pursuits, and a dependence, in this case perhaps a desire to be sheltered by society as represented by the Government and the University. Such a desire for dependence could very conceivably be the outgrowth of the youths having been pushed into the mature role of becoming aggressors for the sake of society, at an age and stage of development where many of them were not ready for such a role.

The Si scale (2) indicates that both groups are generally like average college students in tendencies toward social participation, despite the fact that, as freshmen, they are new to the University, and are, further, students who have demonstrated a felt need for specialized help from the Counseling Center. Students come to this Counseling Center on a purely voluntary basis.

These conclusions cannot be applied to student groups as a whole without reservation, since this study was limited to freshman men; and even these may not be typical of freshman men as a whole, since little is known at present as to what "type" of student seeks out the services of the Counseling Center. It does not seem unlikely, however, that the subjects of the present study are more or less representative of today's student body;

and inasmuch as psychometric tests are, at the present stage of student personnel work, used more frequently on students who come for help to some specialized person or agency than on the entire college population, it is hoped that the present findings may be of some use to those who are attempting to aid students in their adjustment during the post-war period.

REFERENCES

1. Darley, J. G. *Clinical Aspects and Interpretation of the Strong Vocational Interest Blank*. New York: The Psychological Corporation, 1941.
2. Drake, L. E. "A Social I.E. Scale for the Minnesota Multiphasic Personality Inventory." *Journal of Applied Psychology*, XXX (1946), 51-54.
3. Meehl, P. E. and Hathaway, S. R. "The K Factor as a Suppressor Variable in the Minnesota Multiphasic Personality Inventory." *Journal of Applied Psychology*, XXX (1946), 525-564.
- 4 Strong, E. K., Jr. *Vocational Interests of Men and Women*. Stanford University, California: The Stanford University Press, 1943.

AWARD IN STUDENT PERSONNEL RESEARCH

C. GILBERT WRENN

University of Minnesota

NOMINATIONS for the Award in Student Personnel Research may now be submitted to the undersigned members of the Committee on Awards of the Council of Guidance and Personnel Associations. At a meeting in Toronto, Canada, July 9th and 10th, 1949, the Board of Representatives of the Council appointed the Committee on Awards to report at the spring meeting in 1951. The award to be given is not a monetary consideration, but is to be in the form of a statement of recognition by the Board of Representatives of the Council of Guidance and Personnel Associations. It is planned to make announcement of the project or projects selected on Council Day each year and to give publicity concerning the selections through professional journals. It is hoped that such recognition will not only serve to call national attention to significant research already completed, but will stimulate further basic research in the field of student personnel.

Although the Council of Guidance and Personnel Associations is concerned with personnel work and personnel research in industry, business, government, and education, the projects to be considered for the first award are those which were completed within the area of personnel work with students in elementary school, high school, college, and university.

The committee has decided to limit its consideration of research for the first award or awards to studies which were published in some form during the period July 1, 1946, through June 30, 1949. It is recognized that there is much valuable research unpublished as yet or that may never be published, but the inclusion of all unpublished studies would place an unmanageable burden upon the committee. Future committees

may be more inclusive at this point and at the same time cover a more restricted range of time.

It may be necessary to grant two awards, one for research conducted by an individual, another for research conducted by an institution or agency. An Honorable Mention List will also be prepared.

Nominations of studies may be made by any member of the constituent organizations of CGPA, whether the author of the study or not, to any member of the Awards Committee. The committee will depend rather heavily upon such nominations although it may in addition review the literature and supplement the nominations made from the field. Nominations may be made through July 31, 1950.

The research may have been completed by an individual, a group of individuals, or an agency. The individual or individuals concerned need not be members of any constituent organization of CGPA. *The nominations should clearly state the fundamental contribution that the research study has made to student personnel work at any level, together with a statement of the limitations inherent in the research. The nominator should state as fully as possible why he thinks the particular study should be given the award. Wherever possible the nominator should send two or more copies of the research study for examination by the committee.*

It is essential to define what is meant by both "research" and "student personnel work." The committee has adopted the definition of research given in Carter V. Good's *Dictionary of Education*: "Research is the careful unbiased investigation of a problem, based insofar as possible upon demonstrable facts and involving refined distinctions, interpretation, and usually some generalization." The research to be considered may fall in either of two general classifications: studies involving directly any of the personnel services listed below; secondly, educational, psychological, or sociological studies of a more basic nature that contribute fundamentally to a change or development in any of the listed personnel services.

The definition of student personnel work is condensed directly from a statement of the Study Commission of the Council of Guidance and Personnel Associations at the Chicago meeting

in 1949. The services ordinarily to be interpreted as student personnel services at various levels of education are the following:

1. The interpretation of the school to the individual.
2. The maintenance of personnel records and the development of their use.
3. The provision of competent counseling to assist the individual in achieving his best educational, vocational, and personal adjustment.
 - a. This service will have access to psychological testing and such other special diagnostic services.
 - b. This service will give vocational information and will be closely correlated with the placement program.
 - c. This service will supplement the counseling efforts of classroom teachers.
4. Physical and mental health services.
5. Remedial services in such areas as speech, hearing, reading and study habits.
6. Supervision and integration of housing and food services.
7. A program of activities designed to induct the individual into his new life and environment as a member of the school community.
8. The encouragement and supervision of group activities significant to the individual.
9. A program of recreational activities designed to promote lifetime interests and skills appropriate to the individual.
10. The treatment of discipline as a learning experience.
11. Financial or similar aid.
12. Opportunities for securing help through part-time and summer employment.
13. Assistance to the individual in finding appropriate employment when leaving school and later in achieving occupational adjustment and advancement.
14. Enrichment of the life of the individual by providing learning and experiences in the area of spiritual and ethical values.
15. Provision of opportunities for making socially desirable adjustments in relation to the opposite sex.
16. The continuing evaluation of student personnel services in order to make them more effective in the life of the individual.

The members of the Committee on Awards are:

Dr. Mitchell Dreese
 George Washington University, Washington, D. C.
 Dean Clifford Houston
 University of Colorado, Boulder, Colorado

Dr. Warren K. Layton
Detroit Public Schools, Detroit, Mich.

Dean Hilda Threlkeld
University of Louisville, Louisville, Ky.

Dr. C. Gilbert Wrenn, Chairman
University of Minnesota, Minneapolis, Minn.

QUICK ESTIMATION OF MULTIPLE R

WILLIAM LEROY JENKINS

Lehigh University

By the short-cut method described below, the multiple R for a test battery can be estimated in a few minutes with a degree of accuracy sufficient for many practical purposes. Even if a Doolittle solution is finally obtained, the method provides a preliminary estimate and a useful cross-check against serious blunders in computation.

Although intended only as a rough-and-ready approximation, the short-cut has shown so far an astonishing agreement with Doolittle multiple R's. In no case has the difference exceeded .02 and in a set of 20 five-variable problems the mean discrepancy was only .005.

Method with Example

1. Arrange the matrix in descending order of validities. Convert r's to E's using Table 1.

<i>r-matrix</i>					<i>E-matrix</i>				
	Val.	B	C	D		Val.	B	C	D
A	.60	.50	.40	.30	A	20.0	13.4	8.4	4.6
B	.50		.20	.20	B	13.4		2.0	2.0
C	.40			.20	C	8.4			2.0
D	.30				D	4.6			

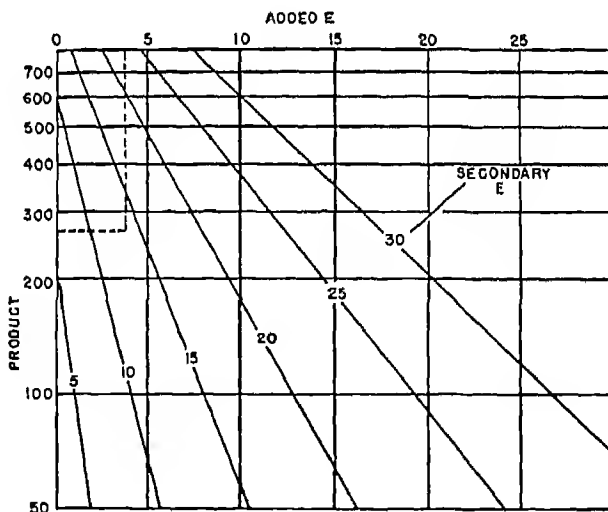
2. Compute the product of the validity E of the first test (Primary) and the intercorrelation E between the first two tests. Find this product on the ordinate scale of Figure 1¹ and move across interpolating between the diagonal lines for the validity E of the second test (Secondary). From this intersection move vertically to the scale of Added E. Add the Primary to the Added E to obtain the multiple E for the first two tests.

Primary	Inter.	Product	Secondary	Added E	Multiple E
20.0	13.4	268	13.4	3.7	23.7 (AB)

¹ The chart in Figure 1 is too small for convenient use. The author will be glad to furnish without charge a photoprint reproduction of the original 8½" x 11" chart on cross-section paper.

TABLE I
Conversion of r to E

r	E	r	E	r	E	r	E	r	E
.10	0.5	.30	4.6	.50	13.4	.70	28.6	.90	56.4
.11	0.6	.31	4.9	.51	13.8	.71	29.6	.91	58.5
.12	0.7	.32	5.3	.52	14.6	.72	30.6	.92	60.8
.13	0.9	.33	5.6	.53	15.2	.73	31.7	.93	63.2
.14	1.0	.34	6.0	.54	15.8	.74	32.7	.94	65.9
.15	1.1	.35	6.3	.55	16.5	.75	33.8	.95	68.8
.16	1.3	.36	6.7	.56	17.2	.76	35.0	.96	72.0
.17	1.5	.37	7.1	.57	17.8	.77	36.2	.97	75.7
.18	1.6	.38	7.5	.58	18.5	.78	37.4	.98	80.1
.19	1.8	.39	7.9	.59	19.3	.79	38.7	.99	85.9
.20	2.0	.40	8.4	.60	20.0	.80	40.0		
.21	2.2	.41	8.8	.61	20.8	.81	41.4		
.22	2.5	.42	9.3	.62	21.5	.82	42.8		
.23	2.7	.43	9.7	.63	22.3	.83	44.2		
.24	2.9	.44	10.2	.64	23.2	.84	45.7		
.25	3.2	.45	10.7	.65	24.0	.85	47.3		
.26	3.4	.46	11.2	.66	24.9	.86	49.0		
.27	3.7	.47	11.7	.67	25.8	.87	50.7		
.28	4.0	.48	12.3	.68	26.7	.88	52.5		
.29	4.3	.49	12.8	.69	27.6	.89	54.4		

FIG. I.
Chart for Added E

The dotted lines in the upper left show the method of finding Added E for step 2 of the problem in the text.

3. Compute the product of the multiple E for the first two tests (Primary) and the *larger* of the intercorrelations of the third test with the first and second. Using this product and the

validity of the third test as Secondary, find the Added E and the new multiple E.

<i>Primary</i>	<i>Inter.</i>	<i>Product</i>	<i>Secondary</i>	<i>Added E</i>	<i>Multiple E</i>
23.7	8.4	199	8.4	1.7	25.4 (ABC)

4. Continue in a similar manner, always using the *largest* of the intercorrelations of the new test with those already forming the multiple.

<i>Primary</i>	<i>Inter.</i>	<i>Product</i>	<i>Secondary</i>	<i>Added E</i>	<i>Multiple E</i>
25.4	4.6	117	4.6	0.6	26.0 (ABCD)

5. Convert the final multiple E to multiple R by reference to Table 1.

Multiple E	26.0
Multiple R	.67 (Doolittle .673)

It will be observed that the process is one of building up the multiple by treating the successive steps as individual three-variable problems, which was the basis of a method² previously published. In the present short-cut, however, the work is considerably reduced, apparently without any serious loss of accuracy.

² Jenkins, W. L. "A Quick Method for Multiple R and Partial r's." (EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT), VI (1946), 273-286.

ERRATUM

In the article by William Leroy Jenkins which appeared in the Spring, 1950, issue of this journal the figure at the bottom of page 143 should be .79 instead of .89.

RECENT PUBLICATIONS RECEIVED

- BROUWER, PAUL J. *Student Personnel Services in General Education*. Washington: American Council on Education, 1949. 317 pp. \$3.50.
- CARTER, HOMER L. J. AND MCGINNIS, DOROTHY J. *Reading Manual and Workbook*. New York: Prentice-Hall, 1949. 120 pp. \$1.75.
- CAVAN, RUTH S.; BURGESS, ERNEST W.; HAVIGHURST, ROBERT J. AND GOLDHAMER, HERBERT. *Personal Adjustment in Old Age*. Chicago: Science Research Associates, 1949. 204 pp. \$2.95.
- CRONBACH, LEE J. *Essentials of Psychological Testing*. New York: Harper & Bros., 1949. 475 pp. \$4.50.
- FREEMAN, FRANK S. *Theory and Practice of Psychological Testing*. New York: Henry Holt & Co., 1950. 518 pp. \$3.50.
- GOODENOUGH, FLORENCE L. *Mental Testing, Its History, Principles and Applications*. New York: Rinehart & Co., 1949. 609 pp. \$5.00.
- GRAY, ROBERT D. AND STAFF. *Selected Personnel Practices of Large Employers in Los Angeles County*. Pasadena: Industrial Relations Section, California Institute of Technology. Circular No. 18. 12 pp. \$1.00.
- GRAY, ROBERT D. AND STAFF. *Survey of Selected Personnel Practices in Los Angeles County*. Pasadena: Industrial Relations Section, California Institute of Technology. Bulletin No. 17. 94 pp. \$2.50.
- HURD, A. W. *Problems of Collegiate Success or Failure with Particular Reference to Professional Schools of Medicine*. Richmond: Bureau of Educational Research and Service, Medical College of Virginia. 124 pp. \$2.50.
- JOHNSON, PALMER O. *Statistical Methods in Research*. New York: Prentice-Hall, 1949. 377 pp.
- LAWRENCE, MERLE. *Studies in Human Behavior. A Laboratory Manual in General Psychology*. Princeton: Princeton University Press, 1949. 184 pp. \$3.50.
- MATHEWSON, ROBERT H. *Guidance Policy and Practice*. New York: Harper & Bros., 1949. 293 pp. \$3.00.
- MOSTELLER, FREDERICK, HYMAN, HERBERT, MCCARTHY, PHILIP J. MARKS, ELI S. AND TRUMAN, DAVID B. *The Pre-Election Polls of 1948*. New York: Social Science Research Council, 1949. 396 pp. \$2.50 (paper), \$3.00 (cloth).
- O'KELLY, LAWRENCE I. *Introduction to Psychopathology*. New York: Prentice-Hall, 1949. 736 pp.
- PARTEN, MILDRED B. *Surveys, Polls and Samples*. New York: Harper & Bros., 1950. 624 pp. \$5.00.

- PEASE, KATHARINE. *Machine Computation of Elementary Statistics*. New York: Chartwell House, 1949. 208 pp.
- PRAY, KENNETH L. M. *Social Work in a Revolutionary Age and other papers*. Philadelphia: University of Pennsylvania Press, 1949. 308 pp. \$4.00.
- REYNOLDS, LLOYD G. AND SHISLER, JOSEPH. *Job Horizons. A Study of Job Satisfaction and Labor Mobility*. New York: Harper & Bros., 1949. 102 pp. \$2.25.
- ROBINSON, VIRGINIA P. *Dynamics of Supervision under Functional Controls*. Philadelphia: University of Pennsylvania Press, 1949. 154 pp. \$2.25.
- SHELDON, WILLIAM H. *Varieties of Delinquent Youth. An Introduction to Constitutional Psychology*. New York: Harper & Bros., 1949. 899 pp. \$8.00.
- SNYGG, DONALD AND COMBS, ARTHUR W. *Individual Behavior: A New Frame of Reference for Psychology*. New York: Harper & Bros., 1949. 386 pp. \$3.50.
- STONE, CALVIN P. *Case Histories in Abnormal Psychology*. Stanford: Stanford University Press, 1949. 106 pp. \$1.75.
- STUIT, DEWEY B., DICKSON, GWENDOLEN S., JORDAN, THOMAS F. AND SCHLOERB, LESTER. *Predicting Success in Professional Schools*. Washington: American Council on Education, 1949. 187 pp. \$3.00.
- SUPER, DONALD E. *Appraising Vocational Fitness By Means of Psychological Tests*. New York: Harper & Bros., 1949. 727 pp. \$6.00. (Text edition, \$5.00.)
- THORNDIKE, ROBERT L. *Personnel Selection: Test and Measurement Techniques*. New York: John Wiley & Sons, 1949. 358 pp. \$4.00.
- TRAVERS, ROBERT M. W. *How to Make Achievement Tests*. New York: The Odyssey Press, 1950. 180 pp. \$2.25.
- WALLIN, J. E. WALLACE. *Children with Mental and Physical Handicaps*. New York: Prentice-Hall, 1949. 549 pp. \$5.00.
- WEITZMAN, ELLIS AND McNAMARA, WALKER J. *Constructing Classroom Examinations: A Guide for Teachers*. Chicago: Science Research Associates, 1949. 153 pp. \$3.00.
- WILLIAMSON, E. G. (Ed.) *Trends in Student Personnel Work*. Minneapolis: University of Minnesota, 1949. 417 pp. \$5.00.
- Manpower Branch, Human Resources Division. Office of Naval Research. *The Development of a Test for Selecting Research Personnel*. Pittsburgh: American Institute for Research, 1950. 33 pp.

THE CONTRIBUTORS

Sidney Adams—Ph.D., University of California, 1933. Job-description writer, U. S. Employment Service, Occupational Research Program, 1935-1937. Research in trade, clinical and other tests, Tennessee Valley Authority, 1937-1946. World War II service in test research, aviation psychology and clinical psychology, 1941-1945. Test developer, U. S. Civil Service Commission, at present.

Robert B. Ammons—Ph.D., University of Iowa, 1946. Assistant Professor of Psychology, Director of Psychological Service for Children, University of Denver, 1946-1948. Assistant Professor of Psychology, Tulane University, 1948-. Dept. of Psychology, University of Louisville, 1949-. Member, American Psychological Association, American Statistical Association, Psychometric Society, Sigma Xi.

P. C. Baker—M.S., Purdue University, 1948. Assistant to the Director, Division of Educational Reference, Purdue University, 1949-

Benjamin Balinsky—Ph.D., New York University, 1940. Psychologist, Bellevue Hospital, 1935-1939. Senior Psychologist, later Head, Psychologist Consultation Service, National Youth Administration, 1939-1942. Civilian Psychological Consultant to War Department, 1942. Psychologist and Counselor, Consultation Service of Vocational Advisory Service, 1942-1947. Evening and summer teaching, City College of New York, 1942-. Instructor, City College of New York, 1947-. Part-time Psychological Consultant to Vocational Services Dept., United Service for New Americans. Author of articles on tests and measurements in vocational and clinical fields. Fellow, American Psychological Association, Division of Counseling and Guidance and Division of Clinical Psychology. Diplomate in Clinical Psychology, American Board of Examiners in Professional Psychology. Member, American Association for the Advancement of Science, American Academy of Political and Social Sciences.

Hubert E. Brogden—Ph.D., University of Illinois, 1939. Instructor, Ohio State University, 1939-1940. Statistician, U. S. Public Health Service, 1940-1942. Personnel Research Section, AGO, 1943-. Author of articles in *Psychometrika*, *Journal of Educational Psychology*, *Psychological Monographs* and *Journal of General Psychology*.

Robert Callis—B.Ed., Southern Illinois Normal University, 1942. With the U. S. Navy, 1942-1946. Counselor, General College, and graduate student, University of Minnesota, 1946.

Raymond B. Cattell—Ph.D., University of London, 1929; D.Sc. (*ibid.*), 1939. Director of City Child Guidance Clinic, Leicester, England, 1932-1937. Research Associate to Professor Thorndike,

Teachers College, Columbia University, 1937-1938. G. Stanley Hall Professor, Clark University, 1939-1941. Lecturer, Harvard University and Civilian Consultant, Adjutant General's Office, 1941-1944. Research Professor in Psychology, University of Illinois, 1945-. Author of *A Guide to Mental Testing*, *Crooked Personalities in Childhood and After*, *General Psychology*, *The Description and Measurement of Personality*, and other books on personality and social psychology, as well as of research articles in American and British journals. Member, American Psychological Association. Fellow, British Psychological Society.

Orria H. Cross—M.S., University of Minnesota, 1945. Teaching Assistant in Psychology, University of Minnesota, 1943-1944. Counselor I, U. S. Employment Service, 1945-1946. Lecturer in Psychology, University of Pittsburgh, 1948-1949. Assistant Professor of Psychology, University of Alabama, 1947- (On leave, 1948-1949). Associate Member, American Psychological Association. Member, Eastern Psychological Association, Southern Society for Philosophy and Psychology, Alabama Academy of Science.

James W. Curtis—M.S., University of Kentucky, 1938. Graduate Assistant, University of Kentucky, 1937-1938. Research Psychologist, United States Forest Service, 1938-1939. Acting Head, Dept. of Psychology, Pikeville College, 1939-1941. Personnel Consultant, Classification Officer, Personnel Liaison Officer, U. S. Army Air Forces, 1941-1947. Supervising Psychologist, Illinois Division of Vocational Rehabilitation, 1947-. Supervising Psychologist, Springfield (Ill.) Mental Hygiene Clinic, 1949. Psychological Consultant to Patton, Evans, Masters Medical Group, 1948-. Author of articles on attitudes, hypnosis, remedial reading, etc., in professional journals. Member, American Psychological Association, Southern Society for Philosophy and Psychology, Illinois Psychological Association, Kentucky Psychological Association, American Association for the Advancement of Science.

N. M. Downie—Ph.D., University of Syracuse, 1948. Instructor in Biology, Robert College, Istanbul, Turkey, 1936-1939. Instructor in Education and Graduate Assistant, Evaluation Service Center, Syracuse University, 1946-1948. Assistant Professor of Education, State College of Washington, 1948-.

Frank M. du Mas—M.A., University of Virginia, 1941. Graduate Student, University of Virginia, 1941-1942. War work and military service, 1942-1945. Instructor in Psychology, University of Denver, 1945-1947. Research Assistant, University of Iowa, 1947-1948. Associate Professor of Psychology, Florida State University, 1948-. On contract, Office of Naval Research, under the guidance of the American Council on Education.

Allen L. Edwards—Ph.D., Northwestern University, 1940. Assistant in Psychology, Ohio State University, 1937-1938. Assistant in Psychology, Northwestern University, 1938-1940. Instructor in Psychology, University of Akron, 1940-1941. Psychologist, Special Study

Group, Military Intelligence, War Dept., 1941-1942. Psychologist, Overseas Branch, Office of War Information, 1942-1943. Assistant Professor of Psychology, University of Maryland, 1943-1944. Consultant, War Relocation Authority, 1943-1944. Associate Professor of Psychology, 1944-1948; Professor of Psychology, 1948-, University of Washington. Author of *Statistical Analysis and Psychology: A First Course in Human Behavior* and author of articles on social psychology, statistics, and scale construction. Fellow, American Psychological Association. Member, American Statistical Association, Biometric Society, Western Psychological Association, Advisory Board, Washington Public Opinion Laboratory. President, State Psychological Association of Washington. Consulting Editor, *Journal of Abnormal and Social Psychology*, *Journal of Applied Psychology*.

Katherine K. Fassett (Mrs. N. C. Fassett)—M.A., University of Wisconsin, 1946. Teaching Assistant, Dept. of Psychology, 1945-1946; Personnel Assistant, Student Counseling Center, Sept.-Dec., 1946; Counselor, Student Counseling Center, 1947-, University of Wisconsin. Associate Member, American Psychological Association, A.P.A. Division of Counseling and Guidance, Association of Midwestern College Psychiatrists and Clinical Psychologists. Member, Midwestern Psychological Association, American Association for the Advancement of Science.

J. P. Guilford—Ph.D., Cornell University, 1927. Instructor in Psychology, University of Illinois, 1926-1927. Assistant Professor of Psychology, University of Kansas, 1927-1928. Associate Professor of Psychology, 1928-1940; Director, Bureau of Instructional Research, 1938-1940, University of Nebraska. Professor of Psychology, University of Southern California, 1940-1942. Director, Psychological Research #3, Santa Ana Air Base, Director, Psychological Research #2, Aviation Cadet Center, San Antonio, Chief, Field Research Unit, Army Air Forces Training Command Headquarters, Fort Worth, Texas; Chief, Dept. of Records and Analysis, Army Air Forces School of Aviation Medicine, Randolph Field, with rank of Colonel, 1942-1946. Professor of Psychology, University of Southern California, 1946-. Fellow, American Association for the Advancement of Science, American Psychological Association. Member, Psychometric Society, Society of Experimental Psychologists, Western Psychological Association, Society of Mathematical Statistics, Southern California Psychological Association.

Arlene B. Heist—M.A., University of Illinois, 1948. Junior Counselor, S.L.A., University of Minnesota, 1948-1950.

Paul A. Heist—M.A., University of Illinois, 1948. Graduate Student, University of Minnesota, at present.

William Leroy Jenkins—Ph.D., University of Michigan, 1936. Instructor, Assistant Professor, Lehigh University, 1935-43. Research Associate, University of California Division of War Research, 1943-44. Supervisor, Training Aids, Columbia University Division of War Research, Submarine Training Section, 1944-45. Associate Professor

of Psychology, 1946-1947; Professor of Psychology, 1947-; Lehigh University. Author of articles on cutaneous sensitivity. Member, American Psychological Association.

C. H. Lawshe - Ph.D., Purdue University, 1940. Member of Faculty, Division of Education and Applied Psychology, 1941-; Professor of Psychology, 1947; Research Associate in Statistical Laboratory, 1948-; Purdue University. Private Consultant to Management, 1942-. Diplomate in Industrial Psychology, American Board of Examiners in Professional Psychology. Fellow, American Psychological Association, American Association for the Advancement of Science. Author, *Principles of Personnel Testing*. Co-author, with Joseph Tiffin and E. J. Asher, of *Workbook for Psychology of Normal People*. Author of articles in professional journals.

William B. Michael -Ph.D., University of Southern California, 1947. Teaching Assistant in Mathematics, 1942-1943; Instructor in Engineering Mathematics, E.S.M.W.T., 1942-1945, California Institute of Technology. Instructor in Mathematics, Pasadena Junior College, 1943-1944. Lecturer in Mathematics, 1944-1947; Lecturer in Education and Psychology, 1945-1947, University of Southern California. Research Associate, College Entrance Examination Board, 1947-1948. Assistant Professor of Psychology, Princeton University, 1947-. Associate Member, American Psychological Association. Member, Mathematical Association of America, Institute of Mathematical Statistics, American Statistical Association, Psychometric Society, Western Psychological Association, Southern California Psychological Association, Phi Beta Kappa, Sigma Xi, Phi Kappa Phi, Phi Delta Kappa.

Adam Poruben, Jr. Ed. D., Columbia University, 1943. Teacher of Science and Mathematics, Saunders Trades School, Yonkers, N. Y., 1934-1944. Personnel Research Technician, Personnel Research Section, AGO, 1944-1945. Research Psychologist, Encyclopedia Britannica Films, 1945-1946. Staff Psychologist, Personnel Division, Metropolitan Life Insurance Co., 1946-. Associate Member, American Psychological Association. Member, American Statistical Association, American Educational Research Association, Committee on Psychological Testing, Commerce and Industry Association of New York.

George Spache—Ph.D., New York University, 1938. Teacher, public schools, N. Y. City, 1930-1936. Psychologist, public and private schools, N. Y. City and Westchester County, 1936-1944. Psychologist, H. L. G. Ledy School, Chappaqua, N. Y., 1944-1948. Consultant, Personnel Services, Rohrer, Hibler and Replogle, Inc., Westchester County, 1948-. Instructor in Education, New York University, 1944-. Instructor, Rutgers University, 1948-. Author of articles on reading, spelling, etc., *The Binocular Reading Test* and *An Incomplete Sentence Test for Use in Industry*. Member, American Psychological Association, Committee on Diagnostic Reading Tests.

Martin Spaggiaro—M.A., New York University, 1949. Served with the U. S. Armed Forces, Military Intelligence, 1942-1945. Intern,

Clinical Psychologist, New York State institutions, 1947-1948. Counselor and Research Assistant, City College Vocational Advisement Unit, 1948- Member, American Psychological Association, Eastern Psychological Association.

Roger G. Stewart—M.A., University of Illinois, 1948. Assistant in Psychology, University of Illinois, 1949-.

Erwin K. Taylor—Ph.D., Northwestern University, 1941. Personnel Examiner, Illinois State Civil Service Commission, 1942-1943. Personnel Research Section, AGO, 1945-. Fellow, American Psychological Association. Member, Psychometric Society, Civil Service Assembly of U. S. and Canada.

Maurice E. Troyer—Ph.D., Ohio State University, 1935. Teacher of biology, athletic coach and Superintendent, Bureau Township Schools, 1923-1929. Assistant Professor of Psychology and Dean of Men, Bluffton College, 1930-1932. Instructor in charge of remedial program, Ohio State University, 1932-1936. Assistant Professor and Associate Professor of Education, Syracuse University, 1936-1940. Associate in Evaluation, Cooperative Study in Teacher Education, American Council on Education, 1940-1943. Professor of Education and Director of Bureau of School Services, Syracuse University, 1943-1945. Director of Evaluation Service Center, 1945-1947, and Director of Psychological Service Center, 1947-1949. Vice-President in charge of curriculum and instruction, Japan International Christian University, 1949-. Author, with Pressey, of *Laboratory Workbook in Educational Psychology*; with Pace, of *Evaluation in Teacher Education*, with Syracuse School of Education Faculty, of *A Functional Program in Teacher Education*, of articles in yearbooks and professional journals. Member, American Psychological Association, American College Personnel Association, American Educational Research Association, American Association for the Advancement of Science, Phi Delta Kappa, Sigma Xi.

G. Gilbert Wrenn—Ph.D., Stanford University, 1932. Vocational Counselor, Stanford University, 1928-1936. Associate Director, General College; Associate Professor, Educational Psychology, 1936-1938; Professor of Educational Psychology, 1938-, University of Minnesota. On military leave, serving in the Bureau of Naval Personnel and Pacific Area as Personnel Officer, 1942-1946. Associate, American Youth Commission, 1939-1941. Consultant, Student Personnel Teacher Education Commission, American Council on Education, 1939-1942. President, National Vocational Guidance Association. Vice-President, Council of Guidance Personnel Association, 1946-. Author and coauthor, *Student Personnel Problems*, *Studying Effectively*, *Aids to Group Guidance*, *Time on Their Hands*, and of professional articles.

Wayne S. Zimmerman—Ph.D., University of Southern California, 1949. Personnel Consultant Assistant, U.S.A.A.F., 1942-1945. Veterans Counselor, 1945-1946; Research Psychologist, 1946-1948, University of Southern California. Psychologist, Sears Roebuck and Company, 1949-. Author of articles in professional journals.



VOLUME TEN, NUMBER THREE, AUTUMN, 1950

<i>A Study of General Education at Syracuse University with Special Attention to the Objectives.</i> N. M. DOWNIE, C. R. PACE AND M. E. TROYER.	359
<i>Educational Growth as Shown by Retests on the Graduate Record Examination.</i> JOSEPH C. HESTON.	367
<i>The Assessment of the Academic Aptitude of the Graduate Student.</i> ROBERT M. W. TRAVERS AND WIMBURN L. WALLACE.	371
<i>Measuring Originality in the Physical Sciences.</i> MILTON M. MANDELL.	380
<i>Probability Approach to Forecasting University Success with Measured Grades as the Criterion.</i> L. J. LINS.	386
<i>Preferences and Behavior Ratings of Dominance.</i> WILLIAM R. BIRGE.	392
<i>Reproducible Scales and the Assumption of Normality.</i> ROBERT G. SMITH, JR.	395
<i>A Factorial Study of Beliefs.</i> J. W. HOLLEY AND CLAUDE E. BUXTON.	400
<i>Opinion and Action: A Study in Validity of Attitude Measurement</i> C. ROBERT PACE.	411
<i>Estimating Intelligence by Interview.</i> JOSEPH V. HANNA.	420
<i>Inclusion of "None of These" Makes Spelling Items More Difficult.</i> MARCIA BOYNTON.	431
<i>A Table and an ABAC for Testing the Significance of Rho.</i> FRANK M. DUMAS.	433
<i>Recent Publications Received.</i>	437
<i>The Contributors.</i>	438

A STUDY OF GENERAL EDUCATION AT SYRACUSE UNIVERSITY WITH SPECIAL ATTENTION TO THE OBJECTIVES

N. M. DOWNIE

State College of Washington

C. R. PACE and M. E. TROYER

Syracuse University

DURING the academic year, 1947-1948, Syracuse University carried out an all-university self-survey. Among the concerns of this survey was a study of the status of the program of general education of the University.

The questions that this study proposed to answer were:

1. What objectives of general education do the students believe to be important?
2. How much of these objectives do the students think that they are achieving in their education at Syracuse?
3. How do the members of the faculty rate the importance of these same objectives and what responsibility does each staff member assume toward helping students achieve these objectives?
4. What is the achievement in general education of Syracuse students as measured by a standardized test of general education?
5. How well-informed are these same students on current events?
6. What are the opinions of the students on some controversial or widely-discussed issues of the day?

This paper will be concerned with the first three of the above questions. The findings of the last three will be reported in later issues of this journal.

Members of the Senior Class, 1948, and of the Sophomore Class, 1950, from the following five colleges of the University participated in the study: Applied Science, Business Administration, Fine Arts, Home Economics and Liberal Arts. An

entire day late in the fall term of 1947 was given over to the testing program involving students. Each student took Form X, the 1947 edition, of the *Cooperative General Culture Test*, the *Time Magazine Current Affairs Test*, reacted to an opinion scale on current issues and to a list of objectives of general education. These same objectives of general education were included in the General Questionnaire completed by each staff member as a part of the survey.

The Objectives of General Education.—A list of objectives of general education developed by a committee of the American Council on Education and reported in *A Design for General Education for Members of the Armed Forces*¹ was modified and used as the basis for this part of the study. This list consists of eighteen items as shown in Table 1.

The student was asked to consider each item in two ways. First, "How important do you consider this knowledge, skill, or understanding as a goal of your education?" Each item was to be marked "very important," "important," "of some importance," "of hardly any importance," or "of no importance." Second, the student was asked to answer the question, "How much are you getting of this knowledge, skill or understanding from college so far?" In this case the answer categories were "much," "some," or "little or nothing."

Each staff member was asked to consider each objective in two ways. The first question was, "How important do you consider this objective as a goal of general education for all students?" This corresponded to the first ratings of the students. The second question was, "What responsibility does your area of instruction assume for helping students make progress toward the attainment of this objective?" Each item was to be marked "direct responsibility," "incidental responsibility," or "outside my area of responsibility." These objectives were rated by 689 faculty members in the various colleges.

The responses of both faculty members and students in rating the importance of these objectives were converted into percentages and the results are shown in Table 1.

The responses of the seniors and the sophomores were first

¹ *Reports of Committees and Conferences*, Series I, Vol. VIII, No. 18. Washington D. C.: The American Council on Education, June, 1944.

compared. For the majority of the items there were no significant differences between the responses of the two classes when

TABLE 1
Ratings of the Importance of the Objectives of General Education by Faculty and Students (Percentages)

Item	Group	1	2	3	4	5
1 Developing good health habits	F	48	30	16	3	1
	S	55	29	11	3	2
2 Understanding the basis of personal and community health	F	38	39	19	1	1
	S	41	37	17	2	2
3 Writing clearly and effectively	F	64	28	6	*	*
	S	57	34	8	1	*
4 Speaking easily and well	F	59	33	7	*	*
	S	75	22	3	*	0
5 Developing social competence and social graces	F	48	37	11	1	*
	S	63	30	6	1	0
6 Understanding other people	F	69	26	4	*	*
	S	85	14	1	0	*
7 Preparing for a satisfactory family and marital adjustment	F	46	34	14	3	2
	S	73	22	4	*	1
8 Discovering personal strengths and weaknesses, abilities and limitations	F	69	24	5	*	1
	S	71	26	3	*	*
9 Understanding world issues and pressing social, political and economic problems	F	57	34	7	1	0
	S	49	39	12	*	*
10 How to participate effectively as a citizen	F	61	32	6	*	*
	S	49	42	7	1	*
11 Understanding scientific developments and processes and their application in society	F	37	44	17	*	*
	S	31	40	25	2	1
12 How to think clearly, meet a problem and follow it to a right conclusion without guidance	F	87	11	1	*	*
	S	85	15	*	0	0
13 Developing an understanding and enjoyment of literature	F	28	45	23	2	1
	S	27	43	26	3	1
14 Developing an understanding and enjoyment of art and music	F	23	38	34	2	2
	S	28	34	33	4	2
15 Understanding the meaning and values in life	F	60	27	9	1	1
	S	65	27	6	1	1
16 Developing a personal philosophy and applying it in daily life	F	55	30	12	1	1
	S	57	30	8	2	2
17 Making a wise vocational choice	F	68	27	4	1	1
	S	88	10	1	*	*
18 Preparing for a vocation	F	59	30	8	1	1
	S	81	16	3	0	*

F—Faculty

S—Students

*—Less than 1 %

1—Very important

2—Important

3—Of some importance

4—Of hardly any importance

5—Of no importance

Chi square was used as a test of significant differences. However, in rating the importance of item 5, "Developing social competence and social graces," the responses of the two classes

were significantly different at the 5 per cent level, with more seniors rating the item "very important." Item 10, "How to participate effectively as a citizen," was also found to be significantly different at the 5 per cent level, again with more seniors rating the objective as "very important." Items 13 and 14, "Developing an understanding and enjoyment of literature" and "Developing an understanding and enjoyment of art and music," were also found to be significantly different at the 5 per cent level. For both of these objectives, more seniors than sophomores rated the items as being "very important."

Considering students' estimates of the amount of each objective they believe they are achieving, item 4, "Speaking easily and well," was significantly different at the 1 per cent level, with the seniors indicating that they were receiving more of this objective than the sophomores. Item 7, "Preparing for a satisfactory family and marital adjustment" was similar to item 4 in all respects. Item 9, "Understanding world issues and pressing social, political and economic problems," was significantly different at the 5 per cent level with the sophomores recording that they were receiving more of this item than the seniors. Item 12, "How to think clearly," and item 14, "Developing an understanding and enjoyment of art and music," were also significantly different at the five per cent level, with the seniors receiving more in both cases.

Thus, in the few instances where there were differences between seniors and sophomores, the seniors tended to regard the objective as more important and to feel that they had made more progress toward its attainment than the sophomores.

The ratings of the faculty and of the students were tested for significant differences, again using Chi square. Twelve of the items were rated differently by the two groups. These are enumerated below. Item 1, "Developing good health habits," was considered to be more important by the students, 5 per cent level. "Writing clearly and effectively," item 3, was rated more important by the faculty, 5 per cent level. Items 4, 5, 6 and 7, "Speaking easily and well," "Developing social competence and social graces," "Understanding other people" and "Preparing for a satisfactory family and marital adjustment," were considered more important by the students, all at the 1 per cent level. Items 9, 10 and 11, "Understanding world

issues and pressing social, political and economic problems," "How to participate effectively as a citizen" and "Understanding scientific developments and processes and their application in society" were all rated as being more important by the staff, all at the 1 per cent level. Items 16, 17 and 18, "Developing a personal philosophy and applying it to daily life," "Making a wise vocational choice" and "Preparing for a vocation" were all considered more important by the students, all at the 1 per cent level.

A comparison of how much of these objectives the students believed they were achieving with the responsibility assumed by the faculty members for the achievement of the objectives is shown in Table 2. A study of this table shows that for most of the objectives, the "Much" column for the students and the "Direct Responsibility" column of the faculty contain the smallest percentages. The exceptions to this over-all pattern are found in item 6, "Understanding other people," where 56 per cent of the students marked that they were receiving "Much" and 33 per cent of the faculty considered this objective to be their "Direct responsibility;" in item 12, "How to think clearly," where 69 per cent of the staff considered this objective to be their direct responsibility and only 37 per cent of the students believed they were receiving much toward the attainment of this objective; and in item 18, "Preparing for a vocation," where 55 per cent of the faculty considered this objective to be their direct responsibility and 48 per cent of the students stated they received much of it.

If we can assume that, in general, there should be some degree of correspondence between the number of faculty members assuming responsibility for an objective and the number of students who feel they are making progress toward its attainment, then we can compare these two ratings. When Chi square was computed for each item, it was found that for all but two items there existed significant differences at the 1 per cent level of confidence. The two for which no differences were found were item 4, "Speaking easily and well," and item 8, "Discovering personal strengths and weaknesses, abilities and limitations." A further analysis of these significant differences showed that, except in items 11, 12 and 17, the students were attaining more of these objectives than the faculty was

assuming responsibility for. This situation might then be an indication of the results of participation in extra-curricular

TABLE 2

Ratings of the Amount of the Objectives of General Education Received by the Students and the Responsibility Assumed by the Faculty for the Achievement of Each Objective (Percentage)

Item	Group	1	2	3
1 Developing good health habits	F	9	27	63
	S	11	45	44
2 Understanding the basis of personal and community health	F	13	28	59
	S	11	45	44
3 Writing clearly and effectively	F	27	54	19
	S	22	51	27
4 Speaking easily and well	F	21	51	28
	S	25	48	27
5 Developing social competence and social graces	F	17	42	41
	S	35	47	18
6 Understanding other people	F	33	39	28
	S	56	36	8
7 Preparing for a satisfactory family and marital adjustment	F	7	21	72
	S	19	36	45
8 Discovering personal strengths and weaknesses, abilities and limitations	F	38	45	17
	S	41	46	13
9 Understanding world issues and pressing social, political and economic problems	F	21	35	44
	S	21	51	28
10 How to participate effectively as a citizen	F	16	41	43
	S	22	55	23
11 Understanding scientific developments and processes and their application in society	F	37	31	32
	S	23	41	36
12 How to think clearly, meet a problem and follow it to a right conclusion without guidance	F	69	22	9
	S	37	50	13
13 Developing an understanding enjoyment of literature	F	15	26	59
	S	21	49	30
14 Developing an understanding and enjoyment of art and music	F	11	22	67
	S	21	30	49
15 Understanding the meaning and values in life	F	22	44	34
	S	25	51	24
16 Developing a personal philosophy and applying it in daily life	F	18	46	36
	S	24	50	26
17 Making a wise vocational choice	F	32	43	25
	S	27	44	29
18 Preparing for a vocation	F	55	29	16
	S	48	42	10

F—Faculty

S—Students

1 For faculty, read "Direct Responsibility"

For students, read "Much"

2 For faculty, read "Incidental Responsibility"

For students, read "Some"

3 For faculty, read "Outside My Area of Responsibility"

For students, read "Little or Nothing"

activities and residence in fraternities and university houses as leading to the realization of some of these objectives of general education.

It should also be borne in mind that, when a student was deciding whether or not he was receiving various amounts of these objectives, he was thinking of no specific courses, but was considering his entire program covering the four or two years that he had been at Syracuse. Each faculty member was considering only his limited area of responsibility. Hence, the differences are actually much greater than indicated by the data because of this difference in the scope of the educational program being rated by the two groups.

The results were next considered in respect to the five different colleges of the University. The ratings of the faculty and students of the various colleges were tested for significant differences using Chi square.

Comparing the results obtained on this check-list of general education objectives within each of the five colleges studied throws some light on the location of responsibility for their attainment and the relative estimates of students' progress toward their accomplishment. For example, the two objectives concerning personal and community health and the one concerning family and marital adjustment were acknowledged as direct responsibilities by a much higher per cent of the Home Economics faculty members than by faculty members in other colleges. Correspondingly, larger numbers of home economics students than students in other colleges felt they were making progress toward these objectives. In contrast, there were no appreciable inter-college differences on the objectives concerned with effective speech and writing. Preparation for effective citizenship and understanding current issues were acknowledged most frequently by faculty and students in Liberal Arts and Business Administration. Except in the College of Fine Arts, almost no faculty members were taking any direct responsibility for helping students understand and enjoy art and music; and almost no students, except in Fine Arts, felt they were achieving much of this objective. These comparisons between colleges are cited as illustrative. Insofar as they are objectives of general education for all students there should probably be a reconsideration of responsibility for their promotion and students in all colleges should feel that they are progressing toward them.

Findings

As a result of having eighteen objectives of general education rated by students and faculty members, the following conclusions can be drawn:

1. The students of Syracuse University consider the attainment of these objectives of general education as important goals of their education.
2. In the curriculum, as it is now organized, the majority of the students feel that they are making "some" but not "much" progress toward the achievement of these goals.
3. The Faculty of Syracuse University considers these same objectives to be important goals of education. However, there is a difference between the importance placed on these various objectives by the faculty and students. The ratings of twelve of these eighteen objectives by the Faculty were significantly different from the students' ratings and eight were rated as being more important by the students.
4. For the achievement of most of these objectives on the part of the students, the majority of the faculty assume no direct responsibility.
5. To the seniors and the sophomores most of these objectives were of equal importance. Four of them, 5, 10, 13 and 14, were rated as being significantly more important by the seniors.
6. For the majority of the objectives, the seniors and sophomores felt that they were receiving about the same amounts, except for items 4, 7, 12 and 14, which the seniors reported to be receiving more of and item 9 which the sophomores received more of.
7. In the five colleges the ratings of the importance of the objectives, the amounts that the students were receiving and the responsibility the faculty members assumed for the achievement of each varied considerably from college to college. The importance placed on an objective and the amount the students received more or less depended on the curriculum of the individual college.

EDUCATIONAL GROWTH AS SHOWN BY RETESTS ON THE GRADUATE RECORD EXAMINATION

JOSEPH C. HESTON
DePauw University

The Problem

EDUCATORS would like to achieve some objective measure of educational growth to demonstrate the progress of students through a university curriculum. One such method of evaluating this growth is offered through the use of The Tests of General Education of the *Graduate Record Examination*. DePauw University is now in a position to make an analysis of the test-retest records of students who took the Examination in 1946 (as sophomores) and repeated the same Examination in 1948 (as seniors). DePauw is one of the universities where sufficient students have been tested and then retested to make such an analysis possible. Even here, however, the present analysis must be restricted to women students, inasmuch as there were not sufficient men sophomores tested in 1946 to make an analysis of men's records worthwhile. Therefore, the present analysis deals with the sophomore versus senior records of 157 DePauw women students.

Results

The most obvious question in this connection would be, how much gain do these students show on the eight Tests of General Education prepared by the Graduate Record Office? In Table 1 will be found the mean score of these students as sophomores on each of the eight tests and again as seniors on each test. It is obvious from the column headed "Mean Gain" that in most cases there was an appreciable gain. Only one area, Physical Science, showed fundamentally no gain at all. This specific result is not entirely unexpected, since very few of these women students took additional physical science courses during their final two years. The greatest gain was exhibited in the area of

Social Studies, followed rather closely by Effectiveness of Expression and the General Education Index, derived from the battery as a whole.

Gains thus exhibited may be taken at their face value, but the question still remains as to their significance. Statistically one approaches this problem by inquiring as to what degree of certainty we know the gain may not have been due to mere chance factors. The solution to this problem is through the use of the critical ratio technique. The critical ratio, found by dividing the difference by the standard error of the difference, may be interpreted as follows: A critical ratio of zero means there are 50 chances in 100 that the gain was due merely to chance. A critical ratio of 1.00 means there are 84 chances in 100 that the

TABLE 1
Critical Ratios of G.R.E. Test Gains for 157 DePaul Women Tested as Sophomores (1946) and Retested as Seniors (1948)

Test	Means		Mean gain	Std. Dev.		Critical Ratio of Gain
	Soph.	Senior		Soph.	Senior	
Math. . .	460	494	34	86.3	97.9	3.26
Phys. Sci..	458	461	3	86.7	91.4	0.30
Biol. Sci..	497	526	29	87.7	94.6	2.82
Soc. Stud.	461	506	45	78.7	88.3	4.77
Literature	491	523	32	81.4	84.3	3.42
Arts . . .	501	528	27	72.1	76.3	3.22
Eff. Exp..	498	538	40	91.6	86.0	3.99
Vocab. . .	453	476	23	75.9	83.9	2.55
G.E. Index.....	469	509	40	81.3	89.4	4.15

true difference is greater than zero; 2.00 means there are 98 chances in 100; while 3.00 can be taken a practical certainty (100 chances in 100).

We find only one critical ratio indicating a difference that is of no consequence, the one for Physical Science, where the mean gain could have been very much a matter of chance. Of the remaining eight critical ratios only two, those for Biological Science and for Vocabulary, are below the 3.00 level, but these two are sufficiently well above the 2.00 level as to mean about 99.5 chances per 100 of being true differences. The critical ratio is not necessarily a measure of the size of the difference, but does indicate if a difference is statistically significant and is not due to chance factors. We may conclude, therefore, that the gains shown on all the tests except Physical Science were sufficiently

appreciable to be well beyond the limits of mere chance and, therefore, represent statistically significant progress from the sophomore year to the senior year. Whether or not this progress is as great as a faculty might wish is obviously still a matter of question.

A second question one would raise in connection with this test-retest program is to what extent did the various tests agree with each other when repeated after two years? This analysis is not to be confused with the concept of statistical reliability.

TABLE 2
Retest Correlations Between Sophomore (1946) and Senior (1948) Administration of G R.E. Tests to 157 DePauw Women

Test	Correlation Soph. vs Senior
Mathematics689
Physical Science.719
Biological Science.616
Social Studies739
Literature639
Arts.752
Effectiveness of Expression705
Vocabulary.851
General Education Index.897

TABLE 3
Correlation Between General Education Index (GRE) and Scholastic Grade Averages (PHR) of 157 DePauw Women

Variables Correlated	Correlation
Soph. GRE vs. Soph PHR	.603
Soph. GRE vs. Senior PHR	.637
Senior GRE vs. Soph. PHR	.549
Senior GRE vs. Senior PHR	.604

Statistical reliability in the process of test construction is determined by test-retest correlation where the examinations are administered relatively close together, so that there is little chance of actual change occurring. However, in this instance the two-year lapse between the tests permitted considerable opportunity for educational gain, not necessarily uniform from student to student on each of the tests. This was due to the situation whereby various students took different curricula and, therefore, made more gains in some of the sub-tests than in some of the others.

In Table 2 we have presented the retest correlations between the sophomore and senior administration of the tests to these same 157 DePauw women. Seven of the sub-tests exhibit retest correlations of .75 or less over the two-year interval. This is not surprising because of the varying degrees of educational growth in each area for each student. Vocabulary did achieve a retest correlation of .85, which indicates considerable consistency from one administration to the next. Vocabulary is not a subject-matter area, but rather an index of general intelligence, and would be expected to show higher retest correlation than the specific subject-matter areas. The General Education Index, exhibiting a correlation of .897, shows that the battery as a whole is remarkably consistent, even when administered with two-year interval of time between test and retest. This high retest correlation for the battery as a whole may be interpreted as meaning students earning a score in the top brackets in the sophomore year would almost certainly earn scores in the top bracket as seniors. In other words, the matter of gain is a relative factor and the degree of gain is marked by considerable consistency throughout the battery as a whole.

A third problem in which one would obviously be interested is the relationship between the GRE General Educational Index and scholastic grade averages at DePauw for these students. In Table 3 we have presented the correlation coefficients between GRE Indexes and grade averages (PHR) for the four possible combinations. Three of these figures are .60 or higher, indicating a strong degree of relationship between GRE scores and university grades. It is interesting to note that the highest correlation is exhibited between GRE index for sophomores and their final senior-grade averages. In this sample at least, it seems it would have been sufficient to give the GRE to sophomores and then to predict final senior-grade averages without recourse to administering the tests again to the seniors. For grade *prediction* this process would have been sufficient, but would not have revealed the *growth* as exhibited in Table 1.

THE ASSESSMENT OF THE ACADEMIC APTITUDE OF THE GRADUATE STUDENT

ROBERT M. W. TRAVERS

and

WIMBURN L. WALLACE

University of Michigan

Introduction

THIS is a study of the assessment of the potentialities of the graduate student and of the criterion of success in graduate school. The study was undertaken at the request of the Executive Board of the Horace H. Rackham School of Graduate Studies of the University of Michigan, and all data were collected within that institution. The study is one of a series of investigations conducted for the administration of the University of Michigan.

Grades as a Criterion of Success in Graduate School

Studies of the prediction of academic achievement are legion, but few of them are particularly concerned with the characteristics of the criterion of academic success. Since it is commonly believed that the failure of tests to make accurate academic predictions is a result of the instability of students' average grades from one semester to the next, it seemed wise to collect some evidence on that point at the beginning of the present study. This was done by finding the correlation between the grades of students in two successive semesters in their field of specialization. From these correlations it is possible to estimate the number of semesters of graduate work that would have to be taken in order for the grade-point average to be a stable criterion of graduate success¹. Table 1 summarizes these data.

The estimated correlation between grades for successive years is based on the Spearman-Brown formula. It is fairly obvious from these data that the average grade for two semesters' work

¹ A stable grade-point average is arbitrarily defined as one that would correlate 0.9 with another grade-point average computed from an equal period of graduate studies.

is a more stable criterion in some areas than in others. It is theoretically quite impossible to predict with any accuracy from test scores or other data the grades which a graduate student of engineering will obtain during a year of graduate studies since grades in that field are highly unstable for a given individual. The stability of average grades in other areas follows that found in previous studies, with the highest in the physical sciences and the lowest in education.

Background of the Present Study

The accurate assessment of the student's potentialities for profiting from work at the graduate level is important for two reasons. First, there is a need for improving selection procedures.

TABLE 1
Stability of Grades

	No. of students	Correlation of Fall and Spring grades	Estimated correlation of grades for 1 year with grades for a second year	Estimated no. of semester's work providing a stable criterion
Social Studies	68	.66	.79	5
Physical Sciences	86	.68	.81	4
Engineering	77	.28	.43	23
Languages and Lit	88	.65	.79	5
Education	68	.53	.69	8

Second, once the student has been admitted it is important to be able to determine how far he should continue graduate work so that he can plan an appropriate program. Every graduate school is familiar with the student who carries a doctoral program to an advanced stage before it is realized that an alternative program would have been a wiser choice. In order to prevent the occurrence of such cases it is necessary to establish a system for appraising the potentialities of the student at an early stage in his career.

There have been two main approaches to the assessment of the graduate student by means of tests. One is that of appraising his "background." In this approach it is assumed that it is most important for the student to enter graduate school with a certain body of information from a variety of subject-matter fields. It makes the additional assumption that a liberal education at

the college level supplies the student with a relatively fixed body of information which can be measured by tests such as the *Graduate Record Examination*. The philosophy of education implied by this approach is more in keeping with the goals of higher education of the last century than with those of the present decade.

The other approach to the assessment of the graduate student is that of determining the extent to which he exhibits the psychological processes and intellectual skills which are important for graduate work. This approach was given limited recognition in the *Graduate Record Examination* in the verbal factor test and is implicit in current proposals for the revision of that examination. It is also illustrated by the use by graduate schools of high-level tests of general ability such as the *C.A.V.D.* scale, and to a lesser extent by the *Miller Analogies Test*.

There are several major reasons at the present time for avoiding the appraisal of the prospective graduate student in terms of his knowledge. First, it is impossible to identify a common body of knowledge which all graduate students should possess, and this will become a progressively more difficult task with the growing emphasis on intellectual skills and arts as major outcomes of a liberal education at the college level. This is true not only insofar as "general background" is concerned but also in the student's field of specialization. Second, it would be most undesirable for graduate schools to indicate that a given body of knowledge was a requirement for graduate work. This would have the evil effect of the graduate schools controlling the undergraduate curriculum in the same way as the colleges have often had the unfortunate role of controlling the curriculum of the secondary school. Third, even if a body of essential knowledge could be identified, there would be no up-to-date examinations for measuring the extent to which this knowledge had been acquired.

For these reasons, the assessment of the graduate student must be largely in terms of the extent to which he has mastery of the intellectual arts and skills necessary for success in graduate work. Following the pattern of the *American Council on Education Psychological Examination* the present investigators devised a test for a higher level of ability which would yield a

linguistic and quantitative score and which would measure processes hypothesized to be important in graduate success. The test will be referred to as the *Academic Aptitude Test, Graduate Level*.

The Nature of Test

All items in the test were multiple-choice with five alternatives. The major portion of the test was liberally timed so as to eliminate the factor of speed. The items were grouped into five parts which are described below:

Part I, Vocabulary.—Eighty words were selected from the technical terminology of the common areas of specialization of graduate work including Physical Sciences, Biological Sciences, Social Studies, Languages, Law, and Philosophy.

Part II, Reading Comprehension.—This test is called a reading test only from custom. It requires the student to reason rather than to memorize what has been read. Questions of the following type which follow some of the passages indicate the kind of mental process which the test involves:

Which one of the following individuals is most likely to have written the above passage?

What is the most important practical oversight in the plan suggested?

What is the main purpose of the author?

What does the author mean by 'regions larger than any empire of antiquity?'

Part III, Verbal Reasoning.—This test involves processes such as the identification of erroneous assumptions, inconsistencies, justifiable in contrast to unjustifiable conclusions, and the making of inferences which are probably but not necessarily correct.

Part IV, Quantitative Reasoning.—This test involves reasoning with numbers, but does not involve mathematics much beyond that taught in junior high school. A few, but not many, of the problems place a considerable emphasis on the ability to understand descriptions of complex data.

Part V, Numerical Ingenuity.—In this test the examinee is presented with a series of numerical problems each of which can be solved by a short method or a long method. The examinee

is instructed to look for the short method of solving each problem. If he does not see the short method at once, he is to pass on to the next problem. This section of the test, unlike the other sections, emphasizes speed.

The original plan of the test was to combine the scores on the first three parts in a verbal-factor score and to add together the scores on the last two parts to provide a numerical reasoning score. The basis for this partition of the test is found in the practice followed by the *American Council on Education Psychological Examination* and in the *Differential Aptitude Battery* both of which provide verbal and numerical scores which have been found to have differential predictive value.

The reliabilities for the verbal and numerical scores calculated by means of the Kuder-Richardson (Formula 21) were found to be 0.86 and 0.86 respectively. The reliability for the total test calculated on the same basis was 0.90. The correlation between the verbal and numerical sections was found to be 0.20. These correlations are based on 484 cases.

The correlation between the verbal and numerical section is much lower than that found with the *American Council on Education Psychological Examination*. In the latter case the correlation between the quantitative and linguistic sections is probably raised considerably by the fact that both sections involve a speed factor.

Validation Procedure

During the academic year 1948-49 the test was administered to 1,111 graduate students. About half of these students were in their first year of graduate work and the remainder had been in graduate school for varying lengths of time. For the purposes of this study, only those students who were registered for 6 or more hours of courses for graduate credit during each semester were included in the investigation. In addition, it seemed desirable to eliminate those foreign students who had taken their undergraduate work in non-English speaking countries. These eliminations reduced to 484 the number of cases included in the study, and these cases were distributed over the various areas of graduate study in the manner shown in Table 2.

Some of these groups are much too small for study; conse-

quently, it seemed advisable to eliminate the biological science, library science and miscellaneous groups from further study

Correlation of Test Scores With Grades

Table 3 summarizes the correlation of test scores with average grades.

Certain facts emerge from this table which throw considerable light on problems of the selection and guidance of graduate students. First, the correlations of test scores with grades in

TABLE 2
Distribution of Students by Field

Field	No. in Each Field
Social Studies	68
Physical Sciences	86
Engineering.....	77
Languages and Literature..	88
Education.....	68
Biological Science.....	28
Library Science	30
Miscellaneous.....	39

TABLE 3
Correlations of Test Scores With Grades

	N	Part					Verbal Score Parts I, II, III	Num. Score Parts IV, V	Total Score
		I Vocab.	II Read.	III Verb. Reas.	IV Num. Reas.	V Inge- nuity			
Social Studies..	68	.46	.24	.09	.10	.05	.36	.09	.31
Physical Science..	86	.08	.46	.31	.33	.09	.18	.27	.27
Engineering. . . .	77	.04	.10	.03	.14	.01	.08	.10	.10
Education.....	68	.45	.42	.38	.16	.28	.49	.24	.47
Lang. and Lit..	88	.41	.27	.34	.35	.24	.47	.37	.50

engineering are negligible in magnitude. This is in accordance with expectations since the criterion represents for practical purposes an unpredictable variable.

Second, there are great differences between the areas of study in the abilities associated with grades. In social studies and languages the part which has the highest predictive value involves vocabulary rather than reasoning. In the physical sciences, on the other hand, it is the sub-tests involving reasoning which have the highest correlation with grades. This does not mean that success in the physical sciences depends upon reason-

ing abilities while success in languages and social studies does not. However, it is consistent with the observation that examinations given in the physical sciences call for problem solving and reasoning abilities while those given in languages and the social sciences commonly call for memory of facts rather than thinking skill. It is probable that the achievement of undergraduates is assigned on the basis of the amount of accumulated knowledge rather than on the basis of the amount of understanding achieved.

TABLE 4
Multiple Correlation Between Test Scores and Average Grades

Field	R
Social Studies49
Physical Sciences52
Engineering17
Education54
Languages and Literature50

TABLE 5
Comparison with Miller Analogies Test

Field	Multiple correlation of Academic Aptitude Test with average grade	Correlation of Miller Analogies Test with average grade
Social Studies49	.18
Physical Sciences52	.38
Engineering17	.09
Education54	.22
Languages and Literature50	.34

Third, the original hypothesis that a verbal and a numerical score represented useful measuring categories does not seem to be consistent with the data. Inspection indicates that differential prediction would be much more effective if the sub-tests were grouped, not into numerical and verbal categories, but into reasoning and vocabulary categories.

Fourth, since there is great variation in the extent to which each of the sub-tests predicts success in each of the areas of study, it would seem that scores on sub-tests should be differentially weighted before they are added together in order to maximize the accuracy of prediction for a given area of study.

Correlations Between Weighted Scores and Average Grades

Table 4 shows the correlations between a composite of sub-scores weighted to give maximum predictions and average grades. These multiple correlations are, with the exception of the one for engineering, of sufficient magnitude to justify the use of the test as one aspect of the assessment and guidance of the graduate student.

It is interesting to compare the above correlations with those found with the *Miller Analogies Test* using a similar criterion of average grade over a year's work in the Horace H. Rackham School of Graduate Studies. This comparison is shown in Table

TABLE 6
Beta Weights for Parts of Test

	Social Studies	Physical Sciences	Engi- neering	Educa- tion	Lang. and Lit.
Vocabulary.....	.50	-.17	-.02	.23	.30
Reading Comprehension.....	.04	.44	.08	.20	.00
Verbal Reasoning.....	-.19	.06	-.05	.22	.17
Numerical Reasoning.....	.06	.23	.16	-.26	.18
Numerical Ingenuity.....	.06	-.09	-.05	.24	.02

TABLE 7
Intercorrelations of Parts of Test

	Part I, Vocab- ulary	Part II Reading	Part III Verbal Reason.	Part IV Num. Reason	Part V Num. Ingenuity
Part I.....		.54	.34	.10	.04
Part II.....			.48	.35	.20
Part III.....				.47	.35
Part IV.....					.64

5. It may be noted that the *Miller Analogies Test* represents a combination of vocabulary and reasoning but it does not permit the differential weighting of these two variables.

Unfortunately, it is not possible to compare the *Academic Aptitude Test* with the results of the *Graduate Record Examination* since the only validation data on the latter is of pre-war vintage, and it is commonly recognized that the predictive value of tests in colleges has for unknown reasons declined in recent years.

The Weights of the Parts in the Composite Scores

It is instructive to examine the weights (beta weights) given to the various parts of the test in the optimum prediction of

grades. These weights indicate the contribution which each part makes independent of all other parts. They are reported in Table 6

This table again suggests rather strongly that in the social studies, education, and languages, grades are based on different criteria than they are in the physical sciences. It indicates also that a test of aptitude for graduate students would be better structured by having a vocabulary and a reasoning section rather than a verbal and a numerical section. In some ways this is rather surprising since the intercorrelations of the sub-tests in the battery show that, in general, the sub-tests cluster into the verbal and numerical categories. This is illustrated in Table 7 which shows the interrelationships between the sub-tests calculated on the basis of 484 cases.

Summary and Conclusions

The present study was concerned with the assessment of the student's aptitude for graduate work. It was demonstrated that if success in graduate work is measured by grades then it is possible only in certain fields to make predictions of success. In the present case, grades in engineering lacked homogeneity from one semester to the next and did not constitute a reasonably predictable criterion. This statement should not be generalized to imply that in other graduate schools grades in engineering would lack consistency from one semester to the next. However, it does imply that graduate schools should from time to time check on the stability of average grades since they are a basis for awarding degrees. If grades are unstable from one semester to the next then any degree awarded on the basis of them is awarded arbitrarily.

The predictive value of a test designed to give a verbal-ability and a numerical-ability score was studied. It was found, however, that the test did not give best predictions when this type of partitioning was used. The evidence indicates that it would be better to partition the test into a vocabulary and a reasoning section and to weight these parts differentially for making predictions in various fields. This finding is important in view of the fact that two of the major testing organizations have announced plans for providing a verbal ability and numerical ability test for the same level of difficulty as the present test.

MEASURING ORIGINALITY IN THE PHYSICAL SCIENCES¹

MILTON M. MANDELL

Examining and Placement Division, United States Civil Service Commission

THE United States Civil Service Commission started in October, 1947, a study of selection methods for physicists, chemists, and engineers. The following report is an interim one which describes the selection methods which seem to predict best the ability to perform research work in the physical sciences, based on a try-out of tests on more than 600 chemists, physicists, and engineers.

The data below are presented in three forms. In the first place, there are presented correlations between test scores and ratings by colleagues and supervisors on a five-point graphic rating scale on an item described as: "Originality of thinking—what is his ability in creative thinking? How original is he in his approach to problems when originality is necessary?" The second method used was to identify those scientists who were engaged in basic research work; this was done in order to determine the correlation between test scores and job performance on the basis of an over-all evaluation on a graphic rating scale by colleagues and supervisors. The third method was to determine the significance of the difference between the mean scores of research personnel and those of non-research personnel on the tests used.

Where the criterion was the summation of ratings of colleagues and supervisors, the method was to add together the ratings by all colleagues and supervisors and to divide these ratings by the total number of ratings, obtaining an average unweighted score.

¹ This study was carried on by the Civil Service Commission as part of its regular program for the improvement of selection methods. Part of the work that was done on this project was performed by persons employed by the American Council on Education in its contract with the Scientific Personnel Division of the Office of Naval Research. Neither organization assumes any responsibility for the contents of this report.

A large number of tests were included in this study. These tests are:

- (1) figure analogies (abstract reasoning)
- (2) Gottschaldt figures
- (3) spatial relations tests, including cube-turning, surface development, and a test developed by a member of the Civil Service Commission which is similar to the block-building test
- (4) formulation
- (5) letter series
- (6) table reading
- (7) vocabulary
- (8) interpretation of data
- (9) hypotheses
- (10) scrambled sentences
- (11) subject matter²

Statistical data are not furnished in this report for many of these tests. In most cases, the reason for the omission of these data is that the correlations were not computed; the scatterplots indicated no significant correlations between the test scores and the criteria. Where the correlations were computed, they were not significantly different from zero.

As will be noted below, many of these tests are quite brief in terms of number of items. This is considered a preliminary study and it was thought advisable to try out a large number of item types. Because of the short testing time available, it was necessary to abbreviate these tests, in some cases probably to a level below that needed for obtaining significant data on their value or lack of value.

1. *Relationship of test scores to ratings on originality.*¹—Subject-matter tests produced significant correlations with ratings on originality. For example, for 35 physicists at the National Bureau of Standards at grades P-1 through P-7, the correlation with a test of approximately 100 items in the basic field of physics was +.59. For 58 chemists at the Eastern Regional Research Laboratory of the Department of Agriculture in grades P-1 through P-4, the correlation with a basic test in

¹ These tests are described in an article, "Selection of Physical Scientists," by Milton M. Mandell and Sidney Adams, *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, VIII (1948) 575-582.

² All correlations included in this report are Pearson product-moment correlations unless otherwise noted.

chemistry of approximately 100 items was $+.46$. For 17 chemists at the Bureau of Standards in grades P-1 through P-6, the correlation was the same, $+.46$. For 53 cases at the Western Regional Laboratory of the Department of Agriculture, the tetrachoric correlation was $+.46$ for the chemistry test, with the sample including chemists at P-1 through P-4. For 19 electronics engineers at the Naval Electronics Laboratory in grade P-2, the correlation between the same basic physics test that was given to physicists and ratings on originality was $+.58$.

In addition to the subject-matter test, other tests produced interesting results. For a test of approximately 35 items prepared by Professor Max Engelhart of the Chicago City Junior Colleges on the ability to evaluate hypotheses, the correlation with ratings on originality for 31 chemists at the Bureau of Standards in P-1 through P-6 was $+.49$. This result did not stand up with the Eastern Regional sample of chemists; however, the correlation for this test at the Eastern Regional Laboratory for 45 chemists engaged in basic research work, when over-all ability in basic research was the criterion, was $+.44$.

In addition to the subject-matter and hypotheses tests described above, a test in basic college mathematics of approximately 30 items correlated $+.41$ for 62 physicists at the Bureau of Standards, with ratings on originality being used as the criterion.

2. *Critical ratios between research and non-research groups.*—A number of tests provided significant differences between the mean scores of those engaged in research work, either basic or applied, and of those engaged in auxiliary work in the sciences, such as testing.

The formulation test, in the form administered at the Bureau of Standards, consisted of 15 items which involved the ability to translate a narrative statement into an algebraic equivalent. It produced significant differences at the 1 per cent level of confidence between 20 chemists engaged in research work and 6 chemists not in research work. The mean score of the research workers was 8.9, and the mean score of the non-research workers was 5.5.

A scrambled sentences test of seven items which involved the ability to determine what the last word of the sentence would be if the sentence were correctly arranged also produced a significant difference at the 1 per cent level of confidence between research and non-research chemists. There were 28 chemists in the research group with a mean score of 2.8, and 8 chemists in the non-research group with a mean score of 1.9.

For physicists, the same formulation test described above also differentiated between research and non-research physicists at the 1 per cent level of confidence. The mean score of 23 research physicists in the formulation test was 11.0, while the mean for 13 non-research physicists was 7.8.

The table reading test of the Air Force was included in the battery of tests. This is essentially a test of carefulness, visual acuity, and attention to detail. For a group of 56 engineers, 17 of whom were in research work and 39 of whom were in non-research work, this test, which takes about 7 minutes to administer, produced a critical ratio significant at the 5 per cent level of confidence. The mean score for research engineers was 17.9; the mean score for non-research engineers was 13.9. These engineers were in the electrical and mechanical fields at the Naval Ordnance Laboratory and were in grades P-1 through P-3.

The same table reading test produced significant results on another population of engineers. This sample of 52 engineers in grade P-3 consisted of 29 research engineers and 23 non-research engineers. In this case, the test score was the number wrong rather than the number right. The mean score of the non-research engineers in terms of number wrong was higher than the mean score of the research engineers, with a difference significant at the 5 per cent level of confidence. The mean score for the research engineers was .57 wrong answers; the mean score of number wrong for the non-research engineers was 1.24 answers.

A similar analysis was based upon 87 engineers at the Naval Electronics Laboratory in grades P-1 through P-4. Thirty-two of these engineers are in research work and 55 are in non-research work. Differences which were significant at the 1 per cent level of confidence, in favor of the research engineers, were

obtained on the formulation test described above and on a vocabulary test. There was a difference of five points in the mean scores on these two tests in favor of the research engineers.

A significant difference at the 1 per cent level of confidence was also obtained on a test of spatial relations in favor of the non-research engineers. This test in spatial relations was prepared by a member of the staff of the United States Civil Service Commission and is similar to the block-building test frequently used. The average score of the non-research engineer on this test was 16 points higher than the average score of the research engineer.

3. *Correlation of test scores with ability in basic research.*—It was possible to obtain a group of 45 chemists from the Eastern Regional Research Laboratory who were engaged in basic research work. These chemists were rated by colleagues and supervisors on a five-point graphic rating scale. The method for determining the rating on basic research ability was to add up the ratings on over-all ability and divide by the number of ratings. In addition to the results with the hypotheses test mentioned above, namely, a correlation of $+.44$, a correlation of $+.61$ was also obtained with the basic chemistry test described above for these 45 chemists in basic research. This was the only group in basic research sufficiently large in numbers to justify the isolation of the group for correlation purposes. A number of other tests were tried out with this group but none produced significant results.

Summary

1. The formulation test seems to have the widest usefulness in differentiating research from non-research personnel. Significant differences at the 1 per cent level were obtained with samples from the fields of physics, chemistry, and engineering.

2. Subject-matter tests also provided pertinent data for physicists, chemists, and engineers, using ratings on originality as the criterion.

3. The other tests produced significant results but their usefulness was more limited. The mathematics test correlated significantly with ratings on originality for physicists; the

scrambled sentences test differentiated between research and non-research chemists; the table reading, vocabulary, and a form of block-building test produced significant data for engineers.

4. The results obtained for these tests from the various samples differed in a number of cases. The differences may be due to differences in the samples, the nature of the work, differences in criteria content, or reliability.

PROBABILITY APPROACH TO FORECASTING UNIVERSITY SUCCESS WITH MEASURED GRADES AS THE CRITERION

L. J. LINS

University of Wisconsin

FOR some time, emphasis has been placed upon the ability to forecast academic success in terms of grade-point averages at the University of Wisconsin. As early as 1909, Dearborn¹ attempted to discover whether relative standings in the secondary school were indicative of academic success at the University. As time progressed, various persons investigated the possibilities of "predicting" grade-point averages through multiple regression.

In September, 1928, 1687 University of Wisconsin freshmen took the *American Council Psychological Examination*. Seven hundred and fifty-six of these freshmen were selected as a sample. All were in attendance at the University for at least one year after taking the examination. American Council percentiles (based upon national norms) and high-school percentile ranks were computed. Zero-order Pearson-Product-Moment Coefficients of Correlation were then calculated between these respective factors and grade-point averages for the freshman year. A multiple coefficient of correlation of .711 resulted.² Thus about 50 per cent of the variance of grade-point average was associated through regression with the two independent variables named. This shows a substantial concomitant variation. Since no factors have been found which would forecast university success better, it was thought advisable to employ the American Council Psychological percentile and the high-school rank percentile in this study.

The approach here employed is one of trying to set up a

¹ Gustav J. Froehlich, "The Prediction of Academic Success at the University of Wisconsin," *The University of Wisconsin Bureau of Guidance and Records, Bulletin* 2574, Series 2358, October 1941, p. 3.

² *Ibid.*, 20-22.

system of success or failure probabilities associated with bivariate quarter ranges of the American Council Psychological and high-school rank percentiles. The sample consists of 1789 freshmen, 1189 men and 600 women, who entered the University of Wisconsin, First Semester, 1948-49. All are residents of Wisconsin and were graduated from Wisconsin high schools. Nine per cent were graduated with secondary classes of 30 or less students, 13 per cent with classes of 31 to 60, 27 per cent with classes of 61 to 150, and 51 per cent with classes of 151 or over.

The *American Council Psychological Examination*, 1947 edition (local norms), and high-school rank percentiles were computed by the Student Counseling Center, University of Wis-

TABLE 1
Mean, Standard Deviation of the Distribution, Standard Error of the Mean, and Critical Ratio of the Difference Between Means for the Samples Used

Variable	Men			Women			C. R. of Diff. of Means
	M	σ_d	σ_M	M	σ_d	σ_M	
Grade-Point Average	1.178	.841	.026	1.409	.810	.035	5.33
High-School Rank	64.936	25.52	.793	77.101	21.30	.920	10.02
Percentile							
American Council							
Psychological							
Percentile	52.315	28.55	.876	47.800	27.129	1.169	3.09

consin. First-semester grade-point averages at the University and the above-mentioned percentiles were recorded. The subjects were divided into two groups by sex in an attempt to discover whether or not the men and women differed significantly in the factors under consideration. If significant differences were found, it would indicate that, for the type of approach herein described, it would be better to forecast university success separately by sex rather than by using the whole group without reference to sex.

The groups are described in Table 1. It is seen that the freshman women maintained a higher mean grade-point average than the men in their first semester at the University of Wisconsin and ranked higher on the average in the high-school classes with which they were graduated. However, the mean

percentile rank of the men on the *American Council Psychological Examination* was higher than that of the women.

Again referring to Table 1, one notes that the means of the men and women on the three variables differ significantly. This would indicate that there is cause for keeping the samples of men and women separate.

A significant association is found between grade-point average and high-school rank and the results of the *American Council Psychological Examination* respectively using the Pearson-Product-Moment method of correlation. The correlation coefficients together with the critical ratios are presented in Table 2.

In setting up the system of success or failure probabilities, each of the two groups, that is men and women, was then subdivided into 16 bi-quarter categories according to percentile

TABLE 2
Coefficients of Correlation with the Dependent Variable of First-Semester Grade-Point Average Together with the Critical Ratio of the Coefficient*

Variable	Men		Women	
	r	C.R. _r	r	C.R. _r
High School Rank Percentile	.58	3.43	.22	1.30
American Council Psychological Percentile41	3.42	.17	1.30

$$* \text{C.R.}_r = \frac{r}{\sigma_r} \text{ where } \sigma_r = \frac{1}{\sqrt{N-1}}$$

rank on the *American Council Psychological Examination* and in high-school class. The resulting "cells" were composed of individuals who had approximately the same percentile ranks. For example, all individuals who ranked between the first and the twenty-fifth percentile on both factors would be in the same "cell."

Each "cell" was then divided according to grade-point averages of the individuals within the "cell." In computing grade-point averages at the University three grade points are assigned for each credit at grade of A, two for each grade of B, one for each grade of C, zero for each grade of D, minus one-half for each condition grade, and minus one for each grade of failure. Averages as computed followed this pattern. Therefore a B average was considered as 2.00-2.99, C as 1.00-1.99, D as 0.00-0.99, and Fail as -1.00-(-0.01). Frequency distributions were then set up for each "cell" and percentages based upon the

total individuals in the "cell" computed. In addition, since a grade-point average of 1.00 is necessary for satisfactory progress in the University, all grade-point averages above 1.00 were considered as successful. As presented in Table 3 and Table 4, this gave the probability of success of entering freshmen.

TABLE 3
*Probability of Academic Success of New Male Freshmen Based Upon High-School Percentile Rank and Percentile Rank American Council Psychological Examination**

High School Rank Percentile	Grade Level	American Council Psychological Percentile							
		0-24		25-49		50-74		75-100	
75-100	B	14	63	19	75	32	83	45	90
	C	49	(49)	56	(107)	51	(136)	45	(228)
	D	33	37	20	25	15	17	9	10
	Fail	4		5		2		1	
50-74	B	6	46	5	55	14	66	14	69
	C	40	(85)	50	(83)	52	(91)	55	(65)
	D	46	54	39	45	27	34	28	31
	Fail	8		6		7		3	
25-49	B	1	29	0	33	8	48	0	64
	C	28	(80)	33	(66)	40	(48)	64	(22)
	D	55	71	41	67	44	52	18	36
	Fail	16		26		8		18	
0-24	B	0	17	6	30	0	47	18	42
	C	17	(60)†	24	(33)	47	(19)	24	(17)
	D	50	83	40	70	32	53	35	58
	Fail	33		30		21		23	

*Probability of success is based upon experience with first-semester freshmen 1948-49 who were graduates of Wisconsin High Schools. The interpretation might be as follows:

It has been our experience that 83 per cent of the men ranking below the twenty-fifth percentile on the *American Council Psychological Examination* (local norms) and in high school class were not successful as first-semester freshmen.

†The number in parentheses is the size of the sample. Numbers above the broken line are probabilities of receiving a C or B or better average. The sum of these two is the probability of success.

In addition to the interpretation as presented in the footnotes of Tables 3 and 4, it seemed desirable to determine a point at which the probability of success would be equal to, or greater than, the probability of failure. Integral values from one to four were assigned to the percentile divisions by quarters of the

American Council Psychological and high-school rank. Thus the quarter 0-24.9 has a value of one, 25-49.9 a value of two, 50-74.9 a value of three, and 75-99.9 a value of four. It is interesting to note for the men, excluding the lowest quarter

TABLE 4
*Probability of Academic Success of New Female Freshmen Based upon High-School Percentile Rank and Percentile Rank American Council Psychological Examination**

High School Rank Percentile	Grade Level	American Council Psychological Percentile							
		0-24		25-49		50-74		75-100	
75-100	B	14	68	19	80	37	90	60	94
	C	54	(63)	61	(107)	53	(106)	34	(115)
	D	27		20		9		6	
	Fail	5	32	0	20	1	10	0	6
50-74	B	2		8		13		22	
	C	38	40	44	52	52	65	50	72
			(47)		(54)		(23)		(14)
	D	49		39		35		14	
	Fail	11	60	9	48	0	35	14	28
25-49	B	0		6		0		0	
	C	44	44	44	50	29	29	0	0
			(25)		(16)		(7)		(0)
	D	36		37		71		0	
	Fail	20	56	13	50	0	71	0	0
0-24	B	0		0		25		0	
	C	31	31	33	33	25	50	0	0
			(16)†		(3)		(4)		(0)
	D	38		67		25		0	
	Fail	31	69	0	67	25	50	0	0

* Probability of success is based upon experience with first-semester freshmen 1948-49 who were graduates of Wisconsin High Schools. The interpretation might be as follows:

It has been our experience that 60 per cent of the women ranking below the twenty-fifth percentile on the *American Council Psychological Examination* (local norms) and between the fiftieth and seventy-fifth percentile in high-school class were not successful as first-semester freshmen.

† The number in parentheses is the size of the sample. Numbers above the broken line are probabilities of receiving a C or B or better average. The sum of these two is the probability of success.

in high-school rank, that if the quarter value of high-school rank is added to the quarter value of the American Council Psychological, generally speaking, a sum of five or more indicates a 50-50 or greater chance of academic success. A sum of six or more indicates at least a 64-36 chance of success and a sum of seven at least 69-31.

The same generally holds true for the sample of women if all "cells" below the fiftieth percentile in high-school rank are eliminated. The exceptions are "cells" 1-3 and 2-2, high-school rank being given first. This may be due to the small frequency and consequent inadequacy of sampling in the lower ranges.

Since percentile rank in high-school class is directly affected by size of class, it is assumed that any forecasts for persons graduated with small classes will not be particularly valid. It was thought advisable, therefore, to either eliminate graduates of small high schools or to arrive at separate success probabilities for this group.

In applying the regression equation in use at the University of Wisconsin for forecasting grade-point averages, it was found that a difference in percentile rank of one would not affect the forecasted grade-point average by more than 0.05 grade point where the size of class is 30 or above. A graduating class of 30 was then selected as the division point between the small and large high school. In applying the same procedure for success probabilities as outlined for the whole group, it was found that eliminating the graduates of small high schools did not affect the probabilities previously reported. It was impossible to arrive at any accurate success probabilities for the small class because of limited size of sampling. Thus the probabilities of the group of less than 30 in graduating class and the group from classes of 31 or more students are not reported here.

It would seem from the results presented that success probabilities could be very beneficial in the educational guidance program both before entering the University and during Freshman Orientation Week. Power of discrimination seems evident. With larger samples and differentiation by colleges, the probability forecast might well take the place of the grade-point average forecast. It might also be more readily understood by the prospective student. Rough measures have been used. It is the feeling of the writer that the results have been interpreted previously as if these rough measures were precision instruments. Therefore possibly too much emphasis has been placed earlier upon the small differences between forecasted grade-point averages.

PREFERENCES AND BEHAVIOR RATINGS OF DOMINANCE

WILLIAM R. BIRGE

Rensselaer Polytechnic Institute

It is well recognized that there is not a necessary correspondence between a person's conduct and his report of his conduct. This situation is generally acknowledged in the field of interest and personality measurement. Meehl and Hathaway (2) have observed that whether or not a person reports his conduct accurately on a questionnaire, his answers may still constitute a significant aspect of his behavior. Kuder (1) points out that there is no necessary relation between scales on his *Preference Record—Personal* and the corresponding areas of actual behavior, but he believes that the use of a number of relatively independent scales is a promising starting point for prediction studies.

This paper, however, is concerned with the question of whether there is a correspondence between conduct and verbal report. The criterion of behavioral ratings was used as the measure of conduct, while the verbal responses were obtained through the use of the *Kuder Preference Record—Personal*.

In connection with another study, the writer obtained sociometric ratings on the trait of dominance from the members of eleven fraternity groups, three sorority groups and two female dormitory groups. Dominant individuals were defined as those who "show the greatest assertiveness and ability to influence others in group situations." The ratings were made on a total of 827 subjects.

With the exception of three small fraternities, the four members from each group who received the highest ratings on dominance and the four members who received the lowest ratings were selected for further study. From the three small fraternities, only two members from each extreme of the domi-

nance ratings were selected. Since, in two groups, three individuals tied for the third from the lowest ratings, there were 58 subjects in the high dominant extreme and 60 subjects in the low dominant extreme.

All of these subjects were requested to fill out the *Kuder Preference Record—Personal*. The response was fairly good. Although 11 members of the high dominant group and 15 members of the low dominant group refused to cooperate in the study, there remained a pool of 92 records for analysis. Forty-seven of these records had been filled out by subjects who received the highest ratings for dominance, while 45 forms had been filled out by subjects who received the lowest ratings for dominance.

TABLE 1

The t's of the Differences Between the Mean Scores on Each of the Six Scales for the High and Low Dominant Groups

Scale	Mean Score (N = 47) High dominant group	Mean Score (N = 45) Low dominant group	t _d	p
A	42.17	38.04	2.09	.04
B	32.40	30.96	.71	.48
C	37.09	33.44	1.54	.12
D	37.92	40.76	1.18	.24
E	52.32	47.16	2.27	.02
H	74.64	81.87	2.10	.04

The 92 records were scored for scales A, B, C, D, and E. The five areas of activity related to these scales are as follows:

- A. Preference for taking the lead and being in the center of activities involving people.
- B. Preference for dealing with concrete problems and everyday affairs rather than interest in imaginative activities.
- C. Preference for thinking, philosophizing, and speculating.
- D. Preference for pleasant and smooth personal relations which are free from conflict.
- E. Preference for activities involving the use of authority and power.

In addition to the five regular scales, the records were also scored on the H scale, an experimental scale designed to measure the degree to which an individual deliberately tries to make a good impression on the test as a whole. It has been

found that individuals who attempt to make a good impression, rather than to answer sincerely, generally receive low H scores. A personal communication from Mrs. Phyllis Cram, of Sears, Roebuck and Co., suggested that the H scale might discriminate between dominant and non-dominant groups. Mrs. Cram tested several administrators with the *Kuder Preference Record--Personal*, and found that the abler administrators tended to receive lower scores on the H scale than did the less able administrators. Mrs. Cram believes that, in this case, there should be no implication that the good administrators were insincere in their answers. She suggests the explanation that these people are "adept at creating a good impression. . . . They are playing their roles expertly, and an effective actor is always sincere even if it is a role."

After the records had been scored, the mean scores on each of the six scales were determined for the high and low dominant groups. The t 's of the differences between these means were then computed. The results of this analysis are presented in Table 1.

As indicated in this table, the differences between the high and low dominant groups on the three scales A, E, and H are significant at the 5 per cent level of confidence. (The H scale means of the two groups were, however, within the "honest" limits.) More specifically, in terms of expressed preferences, these results indicate that the highly dominant person tends to differ from the person with low dominance ratings as follows:

- (1) he prefers to take the lead and be in the center of activities involving people;
- (2) he prefers activities involving the use of authority and power;
- (3) he prefers activities ordinarily chosen by people trying to make a good impression.

REFERENCES

1. Kuder, G. Frederic. *Manual for Kuder Preference Record--Personal*. Chicago: Science Research Associates, 1949.
2. Meehl, Paul E. and Hathaway, Starke R. "The K Factor as a Suppressor Variable in the Minnesota Multiphasic Personality Inventory." *Journal of Applied Psychology*, XXX (1946), 525-564.

REPRODUCIBLE SCALES AND THE ASSUMPTION OF NORMALITY¹

ROBERT G. SMITH, JR.

University of Illinois

THE more commonly used statistical tests of hypotheses assume that the universe of values, as measured, is normally distributed. In some instances, the distribution of scores which an investigator obtains from his sample gives him practically no confidence that this condition is met. That considerable thought has been given to this problem is clear from the recent review by Mueller (6) of numerical transformations. The purpose of the present paper is to examine some of the characteristics of the relatively new technique of reproducible scales from the standpoint of their use with statistics requiring normality assumptions.

The technique of "Scale Analysis," originated by Guttman (2), has attracted considerable attention, since it promises to lead to the construction of tests which are unidimensional. Loevinger (4, 5), in the area of tests of ability, has dealt with the same problem in presenting techniques leading to the construction of "Homogeneous Tests," as she prefers to call them.

Tests are used for two major purposes: to order individuals in the characteristic being measured, and to test hypotheses concerning characteristics. The former may not involve the assumption of normality; the latter requires this assumption if the hypotheses are to be tested with statistical techniques such as the critical ratio, t , and analysis of variance. Some deviation from normality may not affect the precision of the statistical tests to any great degree. If, however, the user of reproducible scales has a distribution of scores which deviates strikingly from normality, then the principle to be described in this paper may permit him to approximate more closely a normal distribution.

¹The writer wishes to express his appreciation to Dr. L. L. McQuitty for his critical comments on this paper.

That the assumption of population normality is not amiss in the case of many reproducible scales gains support from research in other forms of measurement. For instance, Thurstone (7, 8) assumed normality for the purpose of developing scaling techniques for tests of intelligence and for paired comparisons. He was then able to test this assumption experimentally. The testing of the assumption of normality in the area of reproducible scales is a topic for further research.

The common feature of both the Guttman and Loevinger techniques is that they aim to construct tests whose items make perfect discriminations. In the case of a dichotomously scored item, no one who fails the item should make a higher total score than one who passes. With a multiple-response item such as is used in attitude scales, no one giving a lower weighted response should make a higher total score than one who gives a higher weighted response. While the major emphasis in the various techniques for "Scale Analysis" has been in reproducing responses to individual items from the total score, it is possible in a perfectly reproducible scale, since it gives perfect discriminations, to deduce the distribution of the total score from the number of individuals giving each response to each item. Table 1 shows how this can be done.

This means that it will be possible, by the selection of items with properly located cutting points, and by the combination of categories in multiple-response items, to obtain a set of items which, when combined, give a normal distribution of the total score. Such a set of items is shown in Table 2. It will be noted that the characteristic of a perfectly reproducible scale which gives a normally distributed total score is that the scale makes relatively few discriminations between individuals in the center of the range of scores, and progressively more as the extremes are approached. While it is, of course, unlikely that a perfect normal curve will appear in practice, we should be able to approximate normality.

If a sufficiently large pool of items is available, the investigator may select the number of items he intends to use in the scale. Then, if he wants, say, eleven items with cutting points, at one-half sigma units apart, reference to the table of area under the normal curve will give him the proportions desired

TABLE I
*Item Responses and Total Scores of Perfectly Reproducible Test **

Item	% Passing	Test Score
1	88	XXXXXXXXXXXXXXXXXXXX
2	84	XXXXXXXXXXXXXXXXXXXX
3	80	XXXXXXXXXXXXXXXXXXXX
4	70	XXXXXXXXXXXXXXXXXXXX
5	60	XXXXXXXXXXXXXXXXXXXX
6	50	XXXXXXXXXXXXXXXXXXXX

* Each X represents 2% of the cases passing the item. Total scores were obtained by summing X's by columns.

TABLE 2
*A Set of Reproducible Items Which Will Give a Normally Distributed Total Score**

[illegible]

* Each X represents 2% of the cases passing the item. Total scores were obtained by summing X's by columns.

in each item category. If he desires a different number of items, the same procedure may be followed.²

As Guttman (3) has pointed out, the rank order of an individual with regard to a scalable universe of content remains invariant no matter which items are used. Therefore, the selection of items to obtain a normally distributed total score will in no way affect the other valuable properties of the scale. In fact, the placing of restrictions on the location of the cutting points may lead to more efficient scales.

Two scales may have equal reproducibility, but yet have different characteristics as regards the differentiation of individuals. (Compare Tables 1 and 2 in this respect.) It is recognized that normal distributions may not be desirable in all purposes to which tests may be put. However, for a given purpose, if the distribution of cutting points be identical in two tests, equal reproducibility will mean equal efficiency.

A recent use of analysis of variance with a reproducible scale is the study of Gage (1). He, recognizing that the data shown in his Table 15 did not form a normal distribution, was cautious in the interpretation of his results. However, if he had a larger pool of items from which to draw, the selection of items with properly located cutting points could have given him a normal distribution of scores.

According to Guttman (2), one of the advantages of scaling theory is that it does away with "untested and unnecessary hypotheses about normal distributions." Although normality assumptions are not required for scale analysis itself, it may be necessary in some of the uses to which scales are put. Therefore, it is desirable to have a principle to use in achieving normal distributions of total scores on reproducible scales and homogeneous tests.

REFERENCES

1. Gage, N. L. *Scaling and Factorial Design in Opinion Poll Analysis. Studies in Higher Education*, No. 61. Lafayette, Ind.: Purdue Univ., 1947.

² After the present paper was completed, it was brought to the author's attention that this technique had been previously described by G. Hausknecht, in an unpublished article, *A Procedure for Determining a Useful Approximation to an Ideal Scale*. Hausknecht, however, does not bring out the importance of the distribution of cutting points in determining scale efficiency.

2. Guttman, L. "A Basis for Scaling Qualitative Data." *American Sociological Review*, IX (1944), 139-150.
3. Guttman, L. *Questions and Answers About Scale Analysis*. Report D-2, Research Branch, Information and Education Division, Army Service Forces.
4. Loevinger, J. *A Systematic Approach to the Construction and Evaluation of Tests of Ability*. *Psychological Monograph*, Vol. LXI, No. 4, 1947.
5. Loevinger, J. "The Techniques of Homogeneous Tests Compared with Some Aspects of Scale Analysis and Factor Analysis." *Psychological Bulletin*, LXV (1948), 507-529.
6. Mueller, C. G. "Numerical Transformations in the Analysis of Experimental Data." *Psychological Bulletin*, XLVI (1949), 198-223.
7. Thurstone, L. L. "Psychophysical Analysis." *American Journal of Psychology*, XXXIII (1927), 368-389.
8. Thurstone, L. L. "The Unit of Measurement in Educational Scales." *Journal of Educational Psychology*, XVIII (1927), 505-524.

A FACTORIAL STUDY OF BELIEFS¹

J. W. HOLLEY

University of Southern California

and

CLAUDE E. BUXTON

Yale University

THE use of tests of false beliefs is currently popular among teachers of beginning psychology. Such tests serve to stimulate the interest of students in the field of psychology and call attention to many misconceptions and prejudices at the outset. The investigation reported in this paper is concerned with this type of test. Our task was to describe such false beliefs in terms of a limited number of underlying variables, obtained by the method of factor analysis. The results of such an investigation should be of value to the teacher of beginning psychology, for they acquaint him with the dimensions of misconception among students. To students of psychometric techniques, the method alone will be the object of concern.

Statistical background of the study.—In the factor-analysis approach, the investigator may analyze a correlation matrix in which either items or individuals function as variables. The inter-individual correlation method, which has often been adopted in factorial investigations in aesthetics, is known as the "inverted" method of factor analysis. One reason for using it in this currently unstructured field is that there are so many available test items that a matrix of a corresponding order is impractical. This is also true in the domain of lay beliefs about behavior and about psychology. For this reason the "inverted" method of factor analysis, or "Q technique" was employed in our study.

After solving a particular inverted factor-analysis problem, we have as a result, a matrix of common factor loadings. There

¹ The first author wishes to express his appreciation to Professor W. Stephenson, visiting professor at the University of Chicago, for advice regarding the Q technique, particularly in relation to the importance of item difficulty in this method.

is one row for each individual and one column for each common factor. The square of each common-factor loading indicates that portion of the total variance in any particular individual's beliefs which can be attributed to a single factorial source.

† The sum of the common-factor variances for each individual is known as his communality. In this study we may regard this figure as an "index of agreement" between a particular individual and the other individuals represented in the matrix. It indicates the extent to which his evaluations of statements of belief were determined, as were those of the other individuals. The remaining portion of variance ($+1.00$ minus the communality value) could be analyzed further into specific and error variance. The specific variance (estimate of reliability for an individual minus his communality) could be interpreted as an "index of individuality" of beliefs, compared to those of other individuals in the investigation. The error variance ($+1.00$ minus the reliability coefficient for the individual) would represent the remaining portion of variance. In our investigation, however, the reliabilities for the various individuals were not obtained. Therefore, analysis beyond the determination of common-factor variances making up the communality for each individual was impossible.

Procedure.—The second author has constructed, over a period of years, a 100 item true-false test of misconceptions². The items retained, for successive editions, were those which showed some biserial correlation with the total score on the test, were not passed or failed by all subjects, and were worded so that as many correct responses were true as were false. Thus, the typical method of finding all of the items students would fail was not used to build this test. This questionnaire is currently used in the beginning classes on the Evanston campus of Northwestern University.

From a group of 500 test papers, secured in the fall of 1948, 30 were randomly selected. From these 30, 20 papers were finally selected to function as the basic variables of the correlation matrix. (The 10 papers which were eliminated were those

² Some of the items were taken from Valentine (5), some from Garrett and Fisher (1), and some were obtained personally from C.d'A. Gerken of the University of Iowa. A mimeographed copy of this test may be obtained by writing to the second author.

TABLE I
Matrix of Correlations

	INDIVIDUALS																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	.57																			
2	.40	.51																		
3	.33	.35	.61																	
4	.33	.52	.40	.69																
5	.17	.25	.30	.34	.36															
6	.58	.26	.40	.34	.45	.74														
7	.10	.10	.57	.10	.09	.37	.68													
8	.59	.43	.33	.25	.04	.37	.13	.48												
9	.16	.06	.50	.25	.04	.18	.53	.07	.44											
10	.14	.46	.20	.59	.17	.23	.39	.19	.07	.44										
11	.37	.00	.11	.05	.20	.18	.28	.19	.13	.10	.35									
12	.29	.12	.23	.22	.22	.43	.33	.06	.25	.00	.43	.38								
13	.32	.20	.49	.42	.31	.42	.27	.18	.32	.29	.02	.25	.44							
14	.23	.37	.66	.31	.20	.40	.43	.03	.41	.24	.39	.09	.37	.53						
15	.34	.29	.36	.30	.30	.51	.20	.03	.22	.05	.20	.21	.45	.35	.39					
16	.33	.23	.32	.38	.11	.27	.47	.10	.16	.09	.22	.10	.22	.49	.36	.32				
17	.34	.12	.11	.02	.16	.20	.08	.37	.19	.00	.32	.13	.44	.36	.22	.03	.34			
18	.42	.42	.67	.62	.48	.77	.53	.38	.28	.45	.41	.50	.63	.47	.40	.53	.20	.90		
19	.04	.24	.42	.25	.14	.10	.68	.05	.28	.25	.05	.26	.20	.35	.45	.34	.08	.53	.51	
20	.14	.20	.45	.58	.05	.13	.36	.30	.32	.21	.33	.05	.16	.44	.16	.18	.04	.44	.42	.51

with the most extreme scores, i.e., those individuals who either answered almost all or very few of the items correctly. The reason for this final selection was that we wished to avoid cell entries, in the tetrachoric correlations, which were close to zero.)

A matrix of 20 variables was thus obtained from the tetrachoric correlations of the scores of each individual with each of the other nineteen individuals. This correlation matrix, with individuals as variables, is presented in Table 1.

TABLE 2
Centroid Factor Loadings (and Communalities)

Individuals	Factor I	Factor II	Factor III	Factor IV	Communalities
1	.582	-.385	-.029	-.318	.589
2	.497	.120	-.378	-.335	.517
3	.726	.197	.084	.215	.619
4	.608	.338	-.500	-.149	.756
5	.351	-.238	-.268	.294	.338
6	.655	-.494	-.261	.123	.756
7	.646	.351	.308	.274	.710
8	.386	-.267	.076	-.474	.451
9	.416	.286	.419	.065	.435
10	.419	.248	-.279	-.250	.377
11	.394	-.233	.319	-.148	.333
12	.396	-.334	.167	.234	.351
13	.598	-.132	-.108	.125	.402
14	.669	.216	.110	.023	.507
15	.529	-.103	-.106	.267	.373
16	.488	.117	-.068	.146	.278
17	.313	-.305	.312	-.189	.324
18	.943	-.091	-.103	.089	.916
19	.521	.341	.097	.271	.471
20	.496	.368	.207	-.271	.498

From this correlation matrix, four centroid factors (see Table 2) were extracted according to Thurstone's centroid method of factor analysis (2). The reference axes were then rotated following Zimmerman's graphic method (6), so as to minimize the number of zero loadings. The rotations are presented in Table 3, while the final rotated factor loadings are presented in Table 4.

Interpretation of factors.—An important problem in the use of the "Q technique" is determining the meaning of the extracted factors. The rotated factor loadings, by themselves, tell us very little about the nature of the factors, for they merely indicate the rank order of the individuals in regard to these dimensions.

Such a rank order leaves much to be desired in clarifying such meanings.

Stephenson (3) attempted to solve this problem in the case of aesthetic judgments by interrogating the subjects in regard to their preferences and by observing the judgments of those individuals who were highly saturated in one factor. Guilford and Holley (2) employed a system of weighted judgments. They

TABLE 3
Rotation of Centroid Axes

Axes	Degrees	Direction
I & II	46	counter-clockwise
I' & IV	50	counter-clockwise
I'' & III	57	counter-clockwise
I''' & II'	27	clockwise

TABLE 4
Rotated Factor Loadings

Individuals	Factor I	Factor II	Factor III	Factor IV
1	.421	-.044	.556	.318
2	.687	.145	.149	-.017
3	.270	.602	.103	.415
4	.790	.352	-.080	.041
5	.260	-.035	-.111	.507
6	.461	-.091	.216	.699
7	.052	.770	.098	.327
8	.305	-.051	.594	.047
9	-.085	.602	.231	.106
10	.552	.249	.070	-.074
11	.009	.133	.507	.244
12	-.017	.073	.218	.545
13	.346	.202	.136	.471
14	.292	.559	.212	.252
15	.257	.216	.007	.510
16	.272	.346	.007	.289
17	-.022	.026	.526	.213
18	.547	.411	.275	.609
19	.142	.614	-.059	.264
20	.249	.561	.330	-.113

obtained the product of the factor loading of the individual by the rating given by the individual to a particular object. From these scores for the various objects in the aesthetics study reported by these investigators, it was possible to arrange the objects according to the magnitude of these scores for the various factors, and, thus, to name the underlying variables.

In order to determine the nature of the factors extracted in our investigation, biserial correlations were obtained, for each

item, between whether or not the items were passed by the various individuals, and the factor loadings of these individuals. A perfect correlation, then, would be one in which all individuals who missed a particular item also obtained the highest factor loadings. A perfect correlation of an item with the loadings of a particular factor would mean that it measured individual differences maximally in regard to that particular dimension³. That is, it would differentiate, most efficiently, those individuals with high factor loadings from those with low factor loadings. Groups of items which differentiated maximally were used as the basis for naming the factors; that is, we selected clusters of items with the highest correlations and observed the common element among them.

Identification of factors.—In the description of the items below, a positive correlation indicates that the item tended to be passed by those individuals with low factor loadings but missed by those individuals with high factor loadings. In the case of negative correlations, the converse is true. Since our factors represent areas of *misconception*, the positively correlated items are most useful as descriptive of *false* belief, while the negatively correlated items are most useful when contrasted to these. As a convenience, the biserial correlation of an item, together with its scoring key and its level of difficulty as indicated by the number of individuals missing the item, will be presented for each statement which is quoted.

Factor I.—This seems to be a factor of general psychological naïveté. It indicates a lack of technical knowledge about psychology. Those items which describe the factor most clearly are:

"The printing on this page is upside-down on your retina."
(true) $r = +.88$ (5 missed)

³In the Q technique, the factor loading of an individual does not represent the amount of a certain factor present, if the concept of "amount" is defined in terms of expected scores from factorially pure tests. The reason for this is that the Q technique assumes that the means (in this case of misconceptions) are equal to zero and that the variabilities are equal to one for all individuals. The squared factor loading in the Q technique represents that portion of the variance in the misconceptions which the individual has which correlates with the various factors. If then we wanted to know "how much" (as defined by the subsequent scores on a hypothetical "pure factor" test of this dimension), we would have to adjust the squared factor loadings for the amount of misconception, as indicated by their total scores, and for the variability of the individual's scores. This adjustment of variances was not carried out in this particular study, although it should be in subsequent studies of this type. It was felt that the individual differences in the means and variabilities were not sufficiently great to necessitate a reworking of the data.

"Rats, cats, and dogs have the power to reason." (true) $r = +.84$ (15 missed)

"There is little that psychology can do for the *normal* person." (false) $r = +.81$ (2 missed)

Factor II. This seems to be a "knowledge of special terminology" factor. The items which were missed on this factor contained terms whose meanings are not clear to the layman. Items with high correlations are:

"Half the people in this country are below average in intelligence." (true) $r = +.98$ (9 missed)

"The unconscious mind is located just above the roof of the mouth, directly back of the nose." (false) $r = +.82$ (6 missed)

It will be noticed that both of these statements require special knowledge about the *terms* contained in them. The terms "average in intelligence" and "unconscious mind," while familiar to psychologists, do involve a terminology above the level to be expected of the layman.

In contrast to these are the two statements which have the highest negative correlations:

"Cats can see in complete darkness." (false) $r = -.81$ (6 missed)

"A dog can sense impending disaster better than a man." (false) $r = -.81$ (14 missed)

While these last two statements do require a kind of special knowledge, namely that pertaining to perception, there is no problem of terminology here.

Factor III.—This factor appears to be the clearest of the four. It has been labelled "conventional morality." The items with the highest positive correlations are:

"The majority of adult criminals are feeble-minded or very nearly so." (false) $r = +.88$ (4 missed)

"A child is born with a sense of good and evil—this is his conscience." (false) $r = +.80$ (6 missed)

"Being spanked may be pleasurable to a child." (true) $r = +.80$ (7 missed)

"A person who won't look you in the eye is probably untrustworthy." (false) $r = +.78$ (1 missed)

Individuals high in this factor seem to have misconceptions about good and evil. They seem to look upon the *conscience* as

something which is inborn. They appear to regard "bad" *behavior* as being more modifiable through the "will" and intellectual choice of the individual than factual evidence would justify.

Factor IV.—This seems to indicate an "over-evaluation of learning ability," particularly of children. Items with the highest positive correlation are as follows:

"Children memorize much more easily than adults." (false)
 $r = +.98$ (14 missed)

"The average infant would learn to walk two months earlier than he does, if he were given the proper training." (false)
 $r = +.91$ (12 missed)

"The sense organs of touch, in a person with normal vision, are just as sensitive as those in a blind person." (true) $r = +.72$ (14 missed)

It is interesting to note that the item "It is probable that man's instinct to fight is the fundamental cause of wars." (false) had a correlation of $-.71$. (5 missed)

Thus, it is possible to determine the factors of a given area and to carry out item analyses for hundreds of items from data from only a relatively few subjects. We may know how well each item measures each factor, as well as the level of difficulty of each item. For these reasons, this technique is particularly recommended for use in relatively unexplored areas such as aesthetics and ethics, where the investigator is faced with the problem of establishing the principal dimensions from an almost infinitely large number of items. To construct tests in an unexplored domain is costly and time consuming, particularly when the investigator does not know which items to start with. In the method suggested in this paper the investigator starts with every kind of item which he thinks might measure something within the domain being considered. The results give him a rough idea of what the basic dimensions are. He also knows what groups of items are the best measures of these dimensions. He may then start a further analysis of the area, building his tests in the direction of the clusters and using the clusters of selected items as the basis for the selection of similar kinds of items.

To demonstrate this method of screening, eight items were

selected from the original 100. Each factor was represented by two items. Each of these items had a high correlation with the factor it represented, but low correlations with the other factors. Each of these items was correlated with the other items, using tetrachoric coefficients on the basis of the twenty individuals' scores who constituted the twenty variables of the original matrix. This matrix, in which the items are the variables, is presented in Table 5. The variables are grouped according to the

TABLE 5
Intercorrelations of Items

		Items							
		(Factor III)		(Factor IV)		(Factor II)		(Factor I)	
		1	2	3	4	5	6	7	8
(III)	2	+	.75						
	3	-.08	+.10						
(IV)	4	-.22	-.18	+	.55				
	5	-.60	+.08	-.45	-.13				
(II)	6	-.35	-.10	-.75	-.22	+	.75		
	7	-.68	-.45	-.35	.00	-.16	-.36		
(I)	8	-.21	.00	+.60	+.40	-.30	-.85	+	.58

TABLE 6
Centroid Factor Loadings (For Items)

Item	Factor I	Factor II
1	.665	-.753
2	.394	-.604
3	.828	.336
4	.255	.356
5	-.545	-.164
6	-.847	-.486
7	-.212	.678
8	.491	.695

factors which they represent. It is interesting to note at this point that variable 8 is the only one which has a significantly high positive correlation with any factor other than the one it was selected to represent (factor I). It is also of interest to know that this item has a positive biserial correlation with factor IV, while the two items representing factor IV have positive biserial correlations with factor I which are considerably above average.

Thurstone's centroid method of factor analysis was then used to extract two centroid factors (Table 6). The fact that only twenty cases were used in the calculations of the tetra-

choric correlations of the matrix placed a limit upon the number of factors that might have been legitimately extracted without going below the threshold of error variance. It is of interest to note, however, that the pattern on the axes of the two extracted factors consisted of four clusters of items. With the exception of

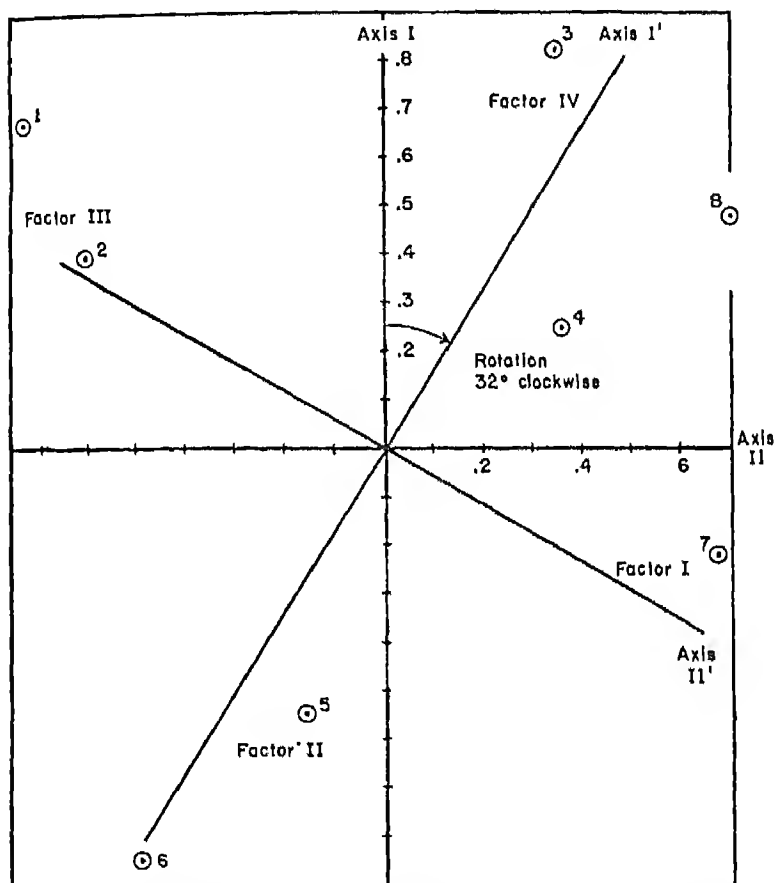


Fig. 1. Projections of Factor Loadings upon Centroid and Rotated Axes.

variable 8, each variable is found with the other member of the pair representative of each factor. Variable 8 had significant loadings on both factor I and factor IV. The projections on the two rotated axes are shown in Figure I.

Summary.—This study was undertaken primarily as a demonstration of methodology, although the factors obtained have

a pedagogical utility. The inverted factor technique was employed so that the extracted factor loadings represented scale values for individuals in regard to these factors. The various items were then correlated with the factor loadings, and the factors were described by those clusters of items which had the highest correlations. The factors which emerged were fairly clear cut. Two items were then selected for each dimension which were highly saturated in that factor. When these eight items were factor analyzed, a pattern of four clusters in two dimensions emerged.

It is suggested that this type of approach be used in unstructured domains in order to obtain a rough idea of the kinds of items which might be used for the further factorial investigation of such areas.

REFERENCES

1. Garrett, H. E. and Fisher, T. F. "Prevalence of Certain Popular Misconceptions." *Journal of Applied Psychology*, X (1926), 411-420.
2. Guilford, J. P. and Holley, J. W. "A Factorial Approach to the Analysis of Variances in Esthetic Judgments." *Journal of Experimental Psychology*, XXXIX (1949), 208-218.
3. Stephenson, W. "The Inverted Factor Technique." *British Journal of Psychology*, XXVI (1935-36), 344-361.
4. Thurstone, L. L. *Multiple Factor Analysis*. Chicago: Univ. of Chicago Press, 1947.
5. Valentine, W. L. "Common Misconceptions of College Students." *Journal of Applied Psychology*, XX (1926), 633-658.
6. Zimmerman, W. "A Simple Method of Orthogonal Rotation of Axes." *Psychometrika*, XVI (1946), 51-55.

OPINION AND ACTION: A STUDY IN VALIDITY OF ATTITUDE MEASUREMENT

C. ROBERT PACE
Syracuse University

THE relationship between opinion and action is a practical topic which has rather basic theoretical importance as well.

Opinion measurement has been attempted by a variety of scientists rather than by a concentration of talent in any single discipline. Thus, we find different techniques employed in public opinion polls, market research, studies of morale, management and job satisfaction, and in education. Political scientists, sociologists, social, clinical, personnel, and educational psychologists, specialists in educational research, and specialists in measurement and evaluation have all made some contribution. While this diversity of approach may be advantageous, it is equally likely that some confusion and superficiality have resulted. Ample documentation of the latter was given in McNemar's (1) critical review of attitude-opinion methodology three years ago. McNemar also stated that relatively few validity studies had been made of attitude and opinion measuring instruments.

Most definitions of attitude accept the proposition that an attitude is a tendency to act for or against some object or value. Most definitions of psychology describe psychology as the science of behavior, or concerned with the prediction and control of behavior. Advances in science are related to the precision of scientific measuring instruments; the value of a measuring instrument is determined in large part by what you can do with the result obtained from it; and what you can do with the result depends on what relationships are known to exist between it and other variables. Thus, the value of an IQ resides largely in the fact that, having it, you can predict a person's behavior or status in quite a variety of circumstances. Likewise, the value of an attitude measurement is largely dependent on knowing

what behavior is associated with it. Opinions are the verbalized expression of attitudes; opinions are not action. But, certainly, some opinions should be correlated with action, just as some aspects of college achievement should be correlated with scores on a college aptitude test.

An opportunity to analyze some data bearing on this problem of opinion validity was provided by the replies of some 2,500 Syracuse University alumni to a sixteen-page questionnaire. (2) (3) This alumni follow-up study was an attempt to describe our educational product rather fully, examining his behavior with respect to some of the major objectives of general education in science, social science, and the humanities. The questionnaire included seven Activity Scales of eleven items each, labelled Politics, Civic Affairs, Religion, Art, Music, Literature, and Science. The subjects checked each activity they had engaged in during the past year. The scales have the property of Guttman-type scales in that participation in the more difficult activities tends to subsume participation in the easier and more common activities. The score on each scale was simply the number of activities checked. Then we had nine Opinion Scales of six items each, labelled Politics, Civic Relations, Government, the World, Philosophy, Art, Music, Literature, and Science. The statements in the opinion scales were written to reflect basic concepts, insights, or appreciations which are among the objectives of general education. Each statement was answered on a five-point scale, from Strongly Agree to Strongly Disagree. Faculty experts in the fields sampled by the opinion scales tended to agree among themselves in their responses to the items and so it was possible to score each scale simply by counting the number of statements on which one's opinion agreed with the opinions of the experts. With only two exceptions, for every statement included in the scales the degree of consensus among answers of the experts exceeded 2 to 1, and for 80 per cent of the items the ratio exceeded 4 to 1. In another section of the questionnaire, we had a list of eighteen objectives of general education, which the alumni rated on a five-point scale of importance, from "very important" to "of no importance." These ratings, of course, are also measures of opinion.

Before reporting correlations between attitudes and activi-

ties, it is appropriate to note the reliability of the scales and the items, for this obviously affects the size of any correlation between them. Six months after our sample of 2,500 had filled out the questionnaire (this represented a 50 per cent return from those who had received it) we sent a second copy of the questionnaire to a small group of 120, receiving 68 in return. The test-retest consistency of scores over this six-months interval was computed, using Pearson product-moment correlations. For the Activity Scales, these ranged from .70 to .89 with a median r of .83. For the nine Opinion Scales, the median correlation was .65, with seven falling between .60 and .70, and two very low ones—.40 and .31. Then we also checked the consistency of responses item by item. For the Activity Scales the

TABLE 1
Correlations Between Scores on Activity Scales and Scores on Opinion Scales
($N = 6,2500$)

Scales	Correlation
Political Activity Score <i>vs</i> Political Opinion Score15
Civic Activity Score <i>vs</i> Civic Opinion Score61
Religious Activity Score <i>vs</i> Philosophy Opinion Score29
Art Activity Score <i>vs</i> Art Opinion Score37
Music Activity Score <i>vs</i> Music Opinion Score40
Literature Activity Score <i>vs</i> Literature Opinion Score33
Science Activity Score <i>vs</i> Science Opinion Score14

average per cent of identical responses was 85, with a range from 83 to 87. For the Opinion Scales the average per cent of identical responses was 75 with a range from 68 to 84.

Correlations between activity and opinion scores are listed in Table 1. The Political Opinion Scale was designed to measure one's belief in the value and importance of individual and group participation in a representative government. The Political Activity Scale is, presumably, a measure of the extent of participation in various political processes, such as discussing and reading about political matters, voting, writing letters, signing petitions, giving and collecting money, etc. One might expect the correlation between two such scales to be considerably higher than .15. The Civic Opinion Scale was intended as a measure of tolerance and acceptance of equality of opportunity for all people. The Civic Activity Scale was intended as a measure of

community participation. The Philosophy Opinion Scale is concerned with acceptance of a Christian and ethical set of values. The Religious Activity Scale is concerned mainly with participation in church-related activities. The opinion scales in Art, Music, and Literature were intended to measure the general sophistication and maturity of understanding in art, music, and literature. The corresponding activity scales were designed to reveal the frequency and depth of engagement in activities related to art, music, and literature.

All these correlations are small, ranging from a low of .01 to a high of .40. We did not construct the scales with the sole thought of correlation between activities and opinions, although we certainly hoped that the people whose opinions reflected the greatest insight and understanding in the various fields would tend also to be most active in those fields. This seems to be true to a limited degree in art, music, literature, and religion, but practically non-existent in politics, civic affairs, and science.

Some of the individual opinion items can appropriately be paired with a corresponding activity item; for other opinions and activities it seemed less reasonable to expect any correspondence. Looking through the questionnaire, I selected 27 opinion statements against which it seemed plausible to compare one or more of 39 activity items. Altogether I had 188 pairs of activity and opinion. For a simple calculation of relationship, I used Thurstone's tables for estimating tetrachoric correlation coefficients. Seventy correlations have been computed and they are the ones which seemed most likely to show some correspondence between opinion and action.

A distribution of the 70 correlations shows a median value of .18, with a fourth at .07 or below, and another fourth at .30 and higher. The lowest was $-.05$. The highest was $+.54$.

All of the correlations above .30 came from the fields of art, music and religion; none came from politics, civic affairs, or science. Literature was not included in these comparisons.

Selected examples of these correlations are shown in Table 2. It is clear that participation in various church-related and other religious activities is definitely correlated with having a favorable opinion toward the significance and importance of religion; but these activities are less clearly related to more general opin-

TABLE 2
Correlations Between Specific Opinions and Specific Actions

Opinions	Actions	Correlation
PHILOSOPHY and RELIGION Disagree with the statement that: Religion has little to offer intel- ligent and scientific people to- day	I belonged to a church	.53
	I contributed a regular sum of money to a church	.43
	I served on some volunteer church committee	.30
	I prayed	.54
	I read selections from my Bible	.24
Rate very important as objective of college education. Understand- ing the meaning and values of life	I belonged to a church	.00
	I contributed03
	I prayed	.15
	I read . . . Bible	.19
ART and MUSIC Disagree with the statement that: Modern painting—impression- ism, expressionism, cubism, sur- realism, and the rest—is mostly the work of crackpots.	I visited an art gallery or mu- seum	.24
	I attended an exhibition of con- temporary painting	.27
	I read one or more books about art, artists, or art history	.38
	I visited an art gallery . .	.31
	I attended an exhibition of con- temporary painting	.45
	I read one or more books about art38
	I listened to some serious music by contemporary composers	.34
	I listened to symphony programs on my radio at least once a month	.37
	I read one or more books about music, musicians, or music his- tory	.37
	I listened to . . . serious contem- porary music35
Disagree with the statement that: The tendency of some modern composers to use strange harmo- nies and discords makes for poor music	I listened to symphony programs20
	I read . . . books about music20
	I listened to . . . serious contem- porary music40
Disagree with the statement that: There has been little or no out- standing music composed in the 20th Century	I listened to symphony programs27
	I read . . . books about music08
	I listened to . . . serious contem- porary music32
Agree with the statement that: Radio should give people much more opportunity to hear good serious music	I listened to symphony programs42
	I subscribed to some orchestral or musical concert series	.33
POLITICS Disagree with the statement that: Sending letters and telegrams to congressmen has little influence on legislation	I wrote a letter or sent a telegram to a public official	.11

TABLE 2—*Continued*

Opinions	Actions	Correlation
POLITICS (continued)		
Agree with statement that: Pressure groups are useful and important features of democratic government	I wrote a letter or sent a telegram to a public official	.10
	I signed a petition for or against some legislation	.06
	I contributed money to some po- litical cause or group	.10
Rate very important as objective of college education: How to par- ticipate effectively as a citizen	I voted in the last primary or lo- cal election	.03
	I signed a petition15
	I wrote a letter or telegram	.10
	I contributed money05
Rate very important as objective of college education: Understand- ing world issues and pressing so- cial, political, and economic prob- lems	I listened at least once a month to speeches and discussion pro- grams on the radio dealing with national and international prob- lems	.26
	I read one or more books about politics	.18

ions about the importance of understanding the meaning and values in life. Opinions about art and music which reflect a sophisticated and mature understanding and interest tend to be accompanied by participation in various art and music activities. In the field of politics, on the other hand, the relations between opinion and action approach zero.

An interesting phenomenon occurs in many of these comparisons between specific opinions and specific actions. People who hold their opinions "strongly" tend to engage in the related activities whether it makes sense or not. For example, among those who feel strongly that sending letters and telegrams has some influence, 37 per cent wrote a letter or sent a telegram. Among those who agree that it has some influence but do not feel strongly about it, 23 per cent wrote a letter or sent a telegram. Among those who had no opinion one way or the other, 10 per cent engaged in the activity. Then, among those who tended to think it had little influence, 18 per cent did it anyway; and among people who were convinced it had little influence, 33 per cent engaged in the activity. Another example: among people who feel strongly that religion does have something to offer intelligent people today, 86 per cent belonged to a church and 76 per cent contributed a regular sum of money to a church.

Among those who agree, but not strongly, that religion is worthwhile today, 72 per cent are church members and 56 per cent contribute money to the church. Among those who have no opinion one way or another, 41 per cent belong to a church and 33 per cent contribute money. But, among those who are convinced that religion has little to offer, 54 per cent belong to a church and 47 per cent contribute money regularly to it. For the activity "I prayed," the percentages drop from 86 to 34 and then rise to 54.

In general, opinions regarding the importance of the various goals of higher education do not exhibit this U-shaped curve in relation to participation in the corresponding activities. For example, among people who rated "Understanding world issues and pressing social, political, and economic problems" as "very important," 82 per cent listened to radio speeches and discussions at least once a month. Among those who rated it as "important," 72 per cent listened. Among people who thought it was "of some importance," 64 per cent listened, and among people who thought it was of little or no importance, 50 per cent listened. With respect to "I read one or more books about politics" the corresponding percentages were 27, 15, 9, and zero.

Or take an illustration from Art. Forty-seven per cent of the people who rated "Developing an understanding and enjoyment of art and music" as "very important" said they had attended an exhibition of contemporary painting. Only 6 per cent of those who considered this to be of little or no importance had attended such an exhibition. Also, in music activities, of those who rated the objective very important, 84 per cent listened to radio symphonies at least once a month in contrast to 46 per cent among those who regarded the objective as of little or no importance.

What conclusions can we draw from these figures? There seems to be some correlation, generally in the .20's and .30's, between belief in the importance of some field and participation in activities in that field. This was true of Art and Music, and to a lesser extent of Religion and Politics. It is also true of science, although I have not reported those correlations. There seems to be a reasonable correlation between specific opinions

and specific actions in Art, Music, and Religion—again generally in the 30's and 40's. In politics, however, I found no correlation higher than $+ .15$ between a specific opinion and a specific action which might be expected to be associated with it.

Many of the correlations in this study may be thought of as rather high. This is so if one considers the probability that there may be an additive or reinforcing effect among related opinions and the further probability that such factors as opportunity for action, multiple actions, and variations in intensity of opinion all may serve to depress the size of correlations between single opinions and single actions. Moreover, tetrachoric coefficients tend to be lower than Pearson product-moment coefficients. The present study is primarily exploratory rather than analytical: it reports relationships in a wide range of fields based on data designed broadly to throw light on the status of the educational product rather than data specifically collected to analyze relationships between opinions and actions. Yet so limited is our knowledge of the validity of many opinion measurements that one of our basic needs is to collect all the information we can from whatever sources so that ultimately critical analysis and theory can be more soundly attempted.

After the failure of the public opinion polls to predict the 1948 Presidential election, attention was focused anew on the relationship between expressed opinion and behavior. The pollsters were quick to claim that their failure in the election had no bearing at all on the value of their regular reports describing the public's attitude on a great variety of complex issues such as labor relations, internationalism, European reconstruction, relations with Russia, etc. The fact is, however, that there is little or no published evidence of the relationship between such attitudes and behavior. Until we have more evidence of the relation between opinion and action, we must regard many of the opinion polls and attitude surveys in the same way that we regard most other magazine and newspaper reports—namely, as interesting observations to be treated with a critical open-mindedness.

Advances in the science of attitude measurement will come in proportion to our ability to establish clear relationships between opinion and action. Until we do this, our so-called meas-

urements will remain purely descriptive. What we must seek is measurement that is both descriptive and predictive of observable behavior.

REFERENCES

1. McNemar, Quinn. "Opinion-Attitude Methodology." *Psychological Bulletin*, XLIII (1946), 289-374.
2. Pace, C. Robert. "Follow-Up Studies of College Graduates." *Growing Points in Educational Research: 1949 Official Report of the American Educational Research Association*, pp. 285-290.
3. Pace, C. Robert. "What Kind of Citizens Do College Graduates Become?" *Journal of General Education*, III (1949), 197-202.

ESTIMATING INTELLIGENCE BY INTERVIEW

JOSEPH V. HANNA

New York University

THE interview had, until recently, been too long neglected among psychologists as yielding promising materials for research. This neglect, in the writer's opinion, is due to two main causes. In the first place, several early and too sketchy experiments yielded results which tended to establish that interviewing techniques and methods were not sufficiently valid to be taken seriously (5, 8, 14). The results of these studies were widely quoted by influential writers, and undoubtedly had the effect of restraining younger clinical and applied psychologists from initiating research projects aiming at the appraisal of interviewing methods and skills. It is a strange paradox that, at the same time, interviewing was nevertheless accepted among psychologists as necessary, and many handbooks and manuals dealing with "acceptable" practices in interviewing were widely used.

A second major reason for the neglect of careful studies of interviewing stems out of the rapid development and use of aptitude tests. Why struggle with a large number of variables in intricate and baffling combination, when a single test which yielded a measurable correlation with a criterion, could be employed? Individuals were selected for specific jobs on the basis of test scores. Intelligence tests were used widely in appraising academic capacity. Yet responsible techniques for dealing with the total person were too frequently absent.

The last few years prior to World War II had witnessed the emergence of a keen interest in a more careful analysis and improvement of interview techniques and skills. Several partially independent efforts contributed to this revival. Greater care was exercised in the interviewing of applicants for employment, and there was developed a more standardized framework for the interview (7). The use of interviewing in adver-

tising research and opinion polling invited more critical attention to such aspects as level of diction, form of the question, and the like. These more objective methods tended to inject themselves into interviewing practice in the areas of clinical and abnormal psychology, vocational counseling, and other fields. Occasional books appeared which epitomized the best research results and applied efforts to interviewing (2, 12). During World War II such instruments as biographical records, rating scales, and careful interview procedures made an impressive contribution to methods of appraising personnel (10). All of these efforts have grown out of the feeling that as valuable as testing is, it is not enough, and that such methods must be supplemented by techniques and procedures which deal with the total person.

The study here reported has to do with the use of interview procedures in estimating the intelligence of clients seeking vocational counsel. By "intelligence" is meant that capacity which is measured more or less accurately by the usual test of intelligence. While the information available to the writer had bearing on a rather wide range of adjustments the information synthesized in the process of the interview is drawn upon only to the extent of indicating the client's cleverness, alertness, or capacity usually referred to as general intelligence.

Procedure

Fifty-four subjects, 50 men and 4 women, were used in the study. They were drawn from applicants to the counseling service of which the writer was in charge, for assistance in deciding what occupation to enter, in choosing appropriate courses of study, and related problems¹. The subjects were taken in order of application, no specifications being made as to age, sex, or other qualities. Care was exercised, however, to eliminate from the sampling all subjects who were introduced to the writer in such a way as to give any indication of background, nature of problem, or abilities and limitations. Those subjects with whom the writer had contacts prior to the pre-

¹ The Personal Counseling Service, West Side Y.M.C.A., New York City. The study was completed shortly before the United States became involved in World War II.

liminary interview were also excluded from the sampling. The estimates of intelligence were based solely on the information secured from the subject, independently of any informal or official reports from other sources. Within the period covered by the study, about 40 per cent of the applicants for the counseling service were eliminated from the sampling due to such prior information and reports.

The subjects were remarkably heterogeneous. They ranged in age from 16 to 44, with a modal age of 17, an average age of 25.9, and a median age of 24.9. Education varied from no formal grade completed to status in graduate and professional school, the average grade completed being 11.6. Intelligence, as measured later, varied from a percentile rank of 2 to 99 plus. The group of fifty men and four women included several refugees from European countries as the result of Nazi persecution.

One of the requirements of the counseling service was that the client fill out *Aids to the Vocational Interview*, an eight-page blank published by the Psychological Corporation. This blank provided space for a fairly comprehensive recording of the client's family background, educational, vocational, and avocational interests and experiences, self-estimates of abilities and the like. It was usually filled out by the client following the preliminary interview. For the clients dealt with in the study, however, the blank was filled out prior to the preliminary interview. The interview required from 20 to 35 minutes. The estimate of intelligence was in all instances limited to the impressions obtained from the subject in the process of the interview. The filled-in *Aids* was helpful, especially, in reducing the time which would have otherwise been required for each interview. Following the interview with each subject the estimate of intelligence was made in terms of a fancied percentile score such as the client would be expected to make on a test of intelligence suitable for entering college freshmen, and in competition with such a selected group. This procedure was decided upon for the sake of uniformity, irrespective of the subject's age or educational background.

The estimate of intelligence was based on the principle of internal consistency, it being assumed that from a reasonably

wide range of cues and impressions there would emerge a constellation or cluster of such items, each item of which is "valid" by agreement with the others. Irrelevant or misleading cues, not being typical of the trend, would be rejected as invalid (1, 11). It should be held in mind that any single item of information offered by the subject, or any single impression of the counselor may or may not be a valid cue. The test of its validity is whether or not it fits in with other cues secured from a variety of sources and directions. If so, then it may be assumed to be valid. It is obvious, however, that the validation of any such cue places a burden upon the interviewer to tap a sufficiently wide area of the subject's background and present status as to reduce to a minimum the chances of error in judgment. If the exploration is too limited in scope any one cue may be weighted unduly, leading to erroneous appraisal of the trait or quality being estimated. Such errors undoubtedly contributed to errors of estimate to be reported later in the present study.

The writer will not attempt to offer a complete list of cues utilized in estimating intelligence. To do so would be impossible due to the subtlety or obscurity of certain cues and relationships synthesized on the basis of overall, intuitive judgment. A listing of the more important and obvious cues, however, may be helpful: (1) subject's report of school grades earned; (2) subject's reported membership in honor clubs and societies; (3) subject's reported standing in school class; (4) reported distinctions and achievement outside of school; (5) reported leadership ability; (6) certain hobbies and activities such as chess, bridge, athletic activities, etc.; (7) conversational ability, use of words, etc.; (8) extent and nature of materials read; (9) activities obviously of compensating nature; (10) range of activities,—varied, or limited; (11) manner and style of responding to questionnaire items; (12) spelling ability; (13) age in relation to grade completed in school,—over-age, accelerated, etc. The following constellation of cues, for example, would point to high intelligence; membership in school honor society, reported high-school average of 95, discriminating use of words in conversation, more interested in English, mathematics, physical sciences and foreign languages, than in the more general

subjects, enjoyment of chess as a hobby, reading of sophisticated books and periodicals. The following constellation would point to limited intelligence; just average grades, "not much of a student," narrow range of vocabulary and lack of discriminating choice of words in conversation, habitual reading of tabloids and popular periodicals, more interested in general

TABLE 1
Age, Grade Completed, Actual and Estimated Percentile Scores, and Errors of Estimation for Fifty-Four Subjects

Age	Grade Comp.	%ile ACE	%ile Ours	%ile Aver.	%ile Est.	Errors of Est.		Age	Grade Comp.	%ile ACE	%ile Ours	%ile Aver.	%ile Est.	Errors of Est.	
						Over Est.	Under Est.							Over Est.	Under Est.
16	11	71	82	76.5	92	15.5		25	14	92	97	94.5	87	7.5	
16	11	70	65	67.5	72	4	5	26	12	50	69	59.5	45		14.5
16	10	10	5	7.5	45	37.5		27	12	93	96	94.5	97	2.5	
16	11	83	—	81*	90	7		28	12	53	88	70.5	60		10.5
16	11	74	59	66.5	86	17.5		28	8	—	1*	50	49		
16	11	61	74	67.5	66		1.5	29	15	88	78	83	65		18
17	11	46	34	40	20		20	29	9	7	—	7*	28	21	
17	11	24	42	33	50	17		30	5	7	4	5.5	6		.5
17	11	96	68	82	82			30	13	78	91	84.5	94		9.5
17	11	44	61	52.5	96	43.5		31	12	99	97	98	60		38
17	10	79	29	54	84	30		31	13	98	66	82	83	1	
17	11	57	71	65	86	21		31	8	99	99	98	92		6
17	11	81	54	68	78	10		31	16	—	90	90*	96	6	
17	11	18	12	15	62	47		31	10	—	32	32*	40	8	
18	11	58	63	60.5	92	31.5		33	8	11	6	8.5	24	15.5	
18	11	55	—	55*	40		15	33	10	54	74	64	45		19
18	13	99	100	99.5	95		4.5	33	18	95	97	96	95		1
18	11	94	89	91.5	47		44.5	35	12	90	—	90*	90		
18	11	52	52	52	55	3		35	12	99	93	96	78		18
19	12	2	33	17.5	45	27.5		35	14	94	—	94*	87		47
20	8	30	62	46	45		1	36	13	76	67	71.5	75	3.5	
20	14	98	88	93	80		13	40	14	91	62	76.5	90	13.5	
20	13	93	84	88.5	83		5.5	40	13	92	91	91.5	72		19.5
21	12	96	95	95.5	65		30.5	41	0	—	54	54*	60	6	
23	15	97	89	93	72		21	43	8	75	95	85	80		5
24	16	84	82	83	87	4		44	16	91	99	95	98	3	
25	16	88	80	84	72		12	—	14	68	—	68*	70	2	

* Where subject took only one test the single score is considered "average."

subjects such as history than in the more exacting subjects, unusual emphasis of physical activities, over-identification with limited hobby. A highly intelligent individual may prefer to read tabloids to other newspapers. The individual with mediocre or low intelligence may unconsciously or otherwise exaggerate his school standing even to the point of indicating honor society membership. Such erroneous cues generally do not fit into the

constellation which seems generally typical of the individual, and can be discarded as invalid.

One further explanation should be made here. The writer made no attempt to weigh or evaluate each cue separately as has been done occasionally in the scoring of interview forms, and application blanks (3, 6, 13, 15). He rather trusted to his judgment to sense the less tangible relationships along with the more obvious cues in arriving at his final estimate.

Following the interview and estimate, all subjects were given a battery of tests including two tests of intelligence,— the *American Psychological Examination for College Freshmen*, and the *Ohio State University Psychological Test*. The first is a time limit and the second a work limit, or power test. Estimated and actual percentile scores and errors of estimate for the 54 subjects are given in Table 1. Distributions of actual percentile scores on the A. C. E. and Ohio, and of estimated percentiles, show the population to be of considerably above-average intelligence. However, the subjects used in the present study are rather typical of clients in general who, throughout the years, applied for counseling to the Personal Counseling Service. All previous studies made of the counselee clientele show above average distributions of intelligence (4).

Results

The actual percentile scores on each test were correlated with the estimated percentiles of intelligence and the two tests were correlated with each other, by the Pearsonian product-moment formula. The following correlations were obtained: A. C. E. with estimates, $r = .71$; Ohio with estimates, $r = .66$; A. C. E. with Ohio, $r = .77$. It will be observed that agreement between estimated percentile scores and scores on each of the two tests of intelligence is just slightly lower than the correlation between the tests. This poses an interesting question as to which of the two instruments or techniques would be the more valid in predicting educational or other achievement.

The results will be examined briefly for the purpose of identifying, if possible, any errors which may account for the deviation of estimates from actual scores. Both tests of intelligence

were taken by 45 of the 54 subjects. For these the average of the two test scores was taken as a basis for comparison with estimated scores. The difference between the estimated score and the average of the test scores, is designated "overestimation" or "underestimation." Examination of Table 1 will show that 15 subjects were overestimated, and 14 were underestimated by a margin of 10 or more percentile points. The average error of overestimation was 15.8 and of underestimation, 14.4 percentile points. The highest error of estimation was 49 percentile points, one subject being overestimated by this margin. One subject was underestimated by a margin of 44 percentile points. For 25, almost half the subjects, however, the error of estimation was 10 or less percentile points.

For those subjects for whom the error of estimate was ten or more percentile points, a study of the records and such notes as had been made following the interview was made with the hope of identifying the factors responsible for the deviation. It is obvious that such an examination cannot be wholly objective. A preliminary inspection, however, had indicated unmistakably the presence of at least one such factor. Examination of data in columns 7 and 8, Table 1, shows clearly the tendency to overestimate the intelligence of younger subjects. Of the 19 subjects eighteen or below, 10 were overestimated by ten or more points, whereas only 3 were underestimated by this margin. Of the 35 who were nineteen or above, only 5 were overestimated by ten or more points, whereas 11 were underestimated by this margin. The tendency to underestimate the intelligence of older subjects, however, is not as clear as the tendency to overestimate the intelligence of younger clients.

The further examination of the filled-in *Aids*, in addition to casting light on the importance of the age factor also indicated roughly several additional factors which seem to have bearing on errors of estimation. These items, impressions, etc., were summarized and appear in Tables 2 and 3. Several items in Table 2 show higher frequency among those overestimated, and in Table 3 for those underestimated.

Reports by subjects of scholarship standing as indicated by grades, position in class, and the like, is apparently the most important single source of errors of estimation, there being a

tendency to rate individuals higher who reported scholarship standing in or near the upper quarter of their class, and a corresponding tendency to rate those lower who reported scholarship below average. Other characteristics of those underestimated were taciturnity, evidence of mediocre reading habits, the early selection of specializing courses such as shop work,

TABLE 2
Characteristics of Subjects Overestimated by Ten or More Percentile Scores

	No.
Reported outstanding specific aptitudes (mathematics, technical, music, art, etc.)	9
Reported high scholarship, regents grades, honors, etc	7
Conversational ability, easy flow of words.	3
Good habits of application.	2
Good looks	2
Miscellaneous	6
(One frequency each for the following characteristics: Well dressed, Practical judgment, Good vocational adjustment, Self-assurance, Foreigner,*—language difficulty, Physical handicap due to birth injury*)	

* In instances such as this it is doubtful if test scores indicate actual level of intelligence

TABLE 3
Characteristics of Subjects Underestimated by Ten or More Percentile Scores

	No.
Mediocre or low scholarship.	13
Taciturn, uncommunicative.	7
Mediocre reading habits	5
Early specializing courses.	5
Emotionally maladjusted	4
Overidentification with narrow interests.	3
Failure to finish courses	2
Miscellaneous	6
(One frequency each for the following characteristics: Poor speller, frequent school absences, poor study habits, marked facial asymmetry, dull appearance, extreme dependence on others)	

typing and the like, and emotional maladjustment; and of those overestimated, good conversational ability, appearance, and positive traits of personality. In a good many cases inflated, sketchy or too modest reports were corrected on the basis of additional items of contra-information. It can readily be seen, however, that a paucity of such "rounding out" information might lead to the acceptance at face value, of questionable

information as fact, resulting in mistakes in estimates. Had the writer been more thorough and searching in his interviewing some of the errors could probably have been avoided or reduced.

Summary of Results

1. It is possible to estimate intelligence test scores with considerable validity on the contents of the interview. Correlations between estimates and test scores are .71 and .66, just slightly lower than the correlation between the two tests, .77.

2. There was a tendency to overestimate the intelligence of younger subjects, and to a lesser extent to underestimate the intelligence of older clients.

3. Underestimation and overestimation of intelligence seem to be related also to reported achievement, reported specific aptitude, negative and positive personality qualities, habits of application, and the like.

Discussion of Results

While it seems clearly possible to estimate intelligence, ability to learn, etc., by interview, it also seems unmistakably clear that the validity of the estimates will depend on two general factors or conditions. First, there must be available a sufficient range of reported information, together with reasonably adequate facilities for interviewing. Second, the experience, competency and skill of the interviewer would seem to be a primary requisite for the validity of estimates.

The relative values of estimates and actual test scores require discussion of a further possibility. Heretofore in the present discussion the differences between actual and estimated scores have been referred to as "errors of estimation" on the traditional assumption that actual test scores should be the more valid in predicting scholastic and related types of achievement. It is obvious, however, that in the absence of objective validation of either estimates or tests for the group of subjects here studied, the relative validities of estimates and tests can only be a matter of conjecture. It seems appropriate to postulate that in dealing with groups such as here reported, careful interviewing based on materials supplied by the individual himself and impressions gained from such interviewing, independently

of official evidence of past performance, grade transcripts, and so on, can be at least as valid in predicting further educational and related achievement as a good test of intelligence. The interview, operating at its highest level, however, is not offered as a substitute for tests of intelligence. The conclusion offered is a reminder to counselors that the storehouse of information available through systematic interviewing, a source too little utilized by many counselors, should not be neglected; and that in the appraisal of the capacities and interests of the client the interview based upon such experience must be regarded as an essential supplement to the more objective measures. In closing it seems appropriate to suggest that counselors in training would find it good practice to utilize interviewing procedures in estimating the intelligence of clients in advance of testing.

REFERENCES

1. Allport, Gordon W. *Personality: A Psychological Interpretation*. New York: Henry Holt and Co., 1937.
2. Bingham, W. V. and Moore, B. V. *How to Interview*. New York: Harper and Brothers, 1931.
3. Goldsmith, Dorothy. "The Use of the Personal History Blank as a Salesmanship Test." *Journal of Applied Psychology*, VI (1922), 149-155.
4. Hanna, Joseph V. "Job Stability and Earning Power of Emotionally Maladjusted as Compared with Emotionally Adjusted Workers." *Journal of Abnormal and Social Psychology*, XXX (1935), 155-163.
5. Hollingworth, H. L. *Vocational Psychology and Character Analysis*. New York: D. Appleton Co., 1929.
6. Hovland, Carl I. and Wonderlic, E. F. "Prediction of Industrial Success from a Standardized Interview." *Journal of Applied Psychology*, XXIII (1939), 537-546.
7. Jenkins, John G. "Characteristics of the Question as Determinants of Dependability." *Journal of Consulting Psychology*, V (1941), 164-169.
8. Magsen, E. H. "How Do We Judge Intelligence?" *British Journal of Psychology*, Supplement No. 9, 1926, 1-108.
9. McMurray, R. N. "Validating the Patterned Interview." *Personnel*, XXIII (1947), 263-272.
10. Newman, Bobbitt and Cameron. "The Reliability of the Interview Method in an Officers Candidate Program." *American Psychologist*, I (1946), 103-109.
11. Primoff, Ernest S. "Correlations and Factor Analysis of the Abilities of the Single Individual." *Journal of General Psychology*, XXVIII (1943), 121-132.

12. Roethlisberger, F. J. and Dickson, W. J. *Management and the Worker*. Cambridge, Mass.: Harvard Univ. Press, 1940.
13. Russell, W. and Cope, G. V. "A Method of Rating the History and Achievement of Applicants for Positions." *Public Personnel Studies*, III (1925), 202-209.
14. Scott, W. D. "Selection of Employees by Means of Quantitative Examinations." *Annals of the American Academy of Political and Social Science*, LXV (1916), 182-193.
15. Snedden, Donald. "Measuring General Intelligence by Interview." *Psychological Clinic*, XVIV (1930), 131-134.

INCLUSION OF "NONE OF THESE" MAKES SPELLING ITEMS MORE DIFFICULT

MARCIA BOYNTON
U. S. Civil Service Commission

A SPECIAL study of the spelling items in its *Clerk and Stenographer-Typist Examinations* has been undertaken by the U. S. Civil Service Commission. All general-test items of these examinations are subjected to systematic statistical evaluation, but further analysis is being made of this one type. The purpose of the study is to determine what elements make for item difficulty, in order to establish guides for improving the control of difficulty in the many alternate forms of examinations required. The amount of information is insufficient as yet to warrant any conclusions. However, a few findings are emerging.

An indication of the value of the alternative "none of these" is one of the preliminary findings. Each of the spelling items has three alternative spellings of a single word, with "none of these" as a fourth alternative. The competitor is instructed to select the correct spelling, if any, or to select the fourth alternative.

Although an item type with only four alternatives is not so desirable as one with five, so few words lend themselves to a sufficient variety of plausible misspellings that the use of five choices was not undertaken. It is recognized that the use of various misspellings is undesirable for the further reason that it emphasizes wrong instead of correct spellings. To avoid both of these objectionable features, each item could include four or five different words. The use of different words in this way, however, presents too great a problem to test constructors in two respects. First, it exhausts the supply of suitable words too quickly in view of the constant need for new sets of examination papers. Second, it increases too greatly the number of words which must not appear in any of the instructions, the vocabulary items, the reading items, or the grammar items of the same test booklet.

The purpose of including "none of these" as an alternative was to increase the number of possible alternatives, thereby reducing the chance that competitors' guesses will be correct.

Analysis of competitors' answers shows that an item that does not include the correct spelling is much more likely to prove difficult than an item in which the correct spelling appears. As is to be expected, an item in which there are two or more points of difficulty is more likely to prove difficult than an item in which there is only one such point. For example, in a sample item, "occasion," a poor speller might wonder whether to use a single consonant, or whether to double both the "c" and the "s". The two findings are consistent, since a constructor would not be able to devise three attractive misspellings of a word unless it contained more than one point of plausible misspelling.

A TABLE AND AN ABAC FOR TESTING THE SIGNIFICANCE OF RHO

FRANK M. DU MAS
University of Texas

I. Introduction

STATISTICIANS have developed several indices of relationship based on ranks. It seems necessary, therefore, to explicitly define the statistical quantity with which this paper is concerned. This statistical quantity derives from Spearman, it is usually called the coefficient of rank difference correlation, and will be referred to in this paper as rho or ρ . Rho is defined as

$$\rho = 1 - \frac{6 \sum d^2}{N(N^2 - 1)}, \quad (1)$$

where, $\sum d^2$ is the sum of the squared differences between paired ranks; N is the number of pairs of ranks.

II. Older Method of Testing the Significance of Rho

The older method of testing the significance of rho is to compute the standard error of rho, divide rho by its standard error, enter the normal probability table with the quotient, and then make a statement concerning the probability of obtaining at least $\rho = 0$ in future samples of the same size taken from the same population. The standard error of rho, $\sigma\rho$, is usually computed from formula (2) as follows:

$$\sigma\rho = \frac{1.04 (1 - \rho^2)}{\sqrt{N - 1}}. \quad (2)$$

There are at least three criticisms of this method of testing the significance of rho. First, formula (2) is only a rough approximation of the standard error of rho. Second, the distribution of rho is markedly skewed when rho is moderate or large and, therefore, the normal probability table should not be used. Third, in those instances where rho is most frequently applied

(say, when $N < 9$), the sampling distributions of rho are most peculiar. When $N = 3$ or 4, they are bimodal; when $N = 5, 6, 7$ or 8, they have a serrated profile. When $N \geq 9$, the distribution may be said to be unimodal and as $N \rightarrow \infty$ the sampling distributions approach the normal distribution as the limit. However, in every case the sampling distributions, for

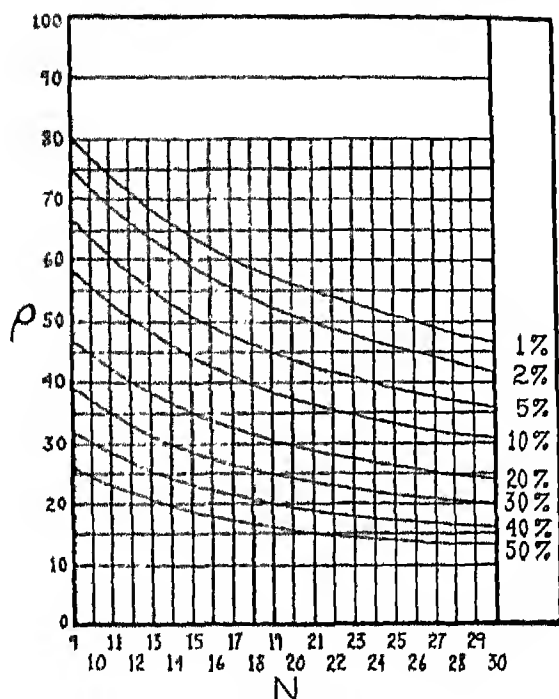


FIG. 1. Abac for Testing the Significance of Rho When $N \geq 9$

$\rho = 0$, are symmetrical. The methods that follow obviate these criticisms to a considerable degree.

III. *Newer Method of Testing the Significance of Rho*

The sampling distributions of rho when $N \geq 9$ ¹ may be said to be unimodal. Actually, these distributions have a saw-tooth profile which tends to smooth out as N increases and approach the normal distribution as the limit. We shall assume the population of rho for samples of $N \geq 9$ to be normally

¹ This value is chosen arbitrarily; we could have chosen 8, 10, 11, 12, etc.

distributed. Under this assumption, we may then enter student's distribution and test the significance of rho. Kendall (1, p. 401) suggests these assumptions and procedures by an

TABLE 1

Table for Testing the Significance of rho when $N < 9$. Values with an Asterisk are Probabilities Rather than Levels of Confidence

N = 4		N = 5		N = 6		N = 7		N = 8	
ρ	% L.C.	ρ	% L.C.	ρ	% L.C.	ρ	% L.C.	ρ	% L.C.
1.00	8	1.00	2	1.00	.0028*	1.00	.0004*	1.00	.000050*
.80	33	.90	8	.94	2	.96	.0028*	.98	.00040*
.60	42	.80	13	.89	3	.93	.0068*	.95	.00114*
.40	75	.70	23	.83	6	.89	1	.93	.0022*
.20	92	.60	35	.77	10	.86	2	.90	.0046*
.00	100	.50	45	.71	14	.82	3	.88	.0072*
		.40	52	.66	18	.79	5	.86	1
		.30	68	.60	24	.75	7	.83	2
		.20	78	.54	30	.71	9	.81	2
		.10	95	.49	36	.68	11	.79	3
		.00	100	.43	42	.64	14	.76	4
				.37	50	.61	17	.74	5
				.31	56	.57	20	.71	6
				.26	66	.54	24	.69	7
				.20	71	.50	27	.67	8
				.14	80	.46	30	.64	10
				.09	92	.43	35	.62	11
				.00	100	.39	40	.60	13
						.36	44	.57	15
						.32	50	.55	17
						.29	56	.52	20
						.25	59	.50	22
						.21	66	.48	24
						.18	71	.45	27
						.14	78	.43	30
						.11	84	.40	33
						.07	91	.38	36
						.04	96	.36	39
						.00	100	.33	43
								.31	46
								.29	50
								.26	54
								.24	58
								.21	62
								.19	66
								.17	70
								.14	75
								.12	79
								.10	84
								.07	88
								.05	93
								.02	98
								.00	100

example in which N equaled 10. We shall use this procedure when $N \geq 9$.

Figure I is an abac to be used in testing the significance of

rho when $N \geq 9$. Formula (3) has been suggested by Kendall (1, p. 401) as appropriate for testing the significance of rho.

$$t = \rho \left(\frac{N-2}{1-\rho^2} \right)^{\frac{1}{2}}. \quad (3)$$

Since $N-2$ are the degrees of freedom, df , we may substitute and solve for ρ . When this is done we have

$$\rho = \sqrt{\frac{t^2}{t^2 + df}}. \quad (4)$$

It was then an easy matter to enter formula (4) with t and df for the various levels of confidence shown in Figure I. The contours in Figure I indicate changes in rho as a function of the sample size with the level of confidence for rejecting the null hypothesis as the parameter.

Figure I may be used in the following manner. Assume a sample of $N = 27$, and $\rho = .38$. Entering Figure I with these values we find that we may reject the null hypothesis at the 5 per cent level of confidence.

Because of the unusual characteristics of the sampling distributions of rho when $N < 9$, the t test of significance would be inappropriate. But it is precisely for the small values of N that a significance test is needed so badly. For example, clinical research is often an intensive study of a few individuals and rho is often used in such situations. Table 1 was constructed with these considerations in mind. Kendall (1, Table 16.2) has tabled the probability of obtaining the various values of Σd^2 for several different values of N . The transformation of Σd^2 and probabilities into rho and levels of confidence is obvious. Table 1 allows us to make (within rounding errors) an 'exact' test of the null hypothesis for samples of 4 to 8 cases.

Table 1 may be used in the following manner. Assume a sample of $N = 7$ and $\rho = .57$. We may then reject the null hypothesis at the 20 per cent level of confidence.

In both Figure I and Table 1 we are testing hypotheses concerning the *absolute* value of rho.

REFERENCES

1. Kendall, M. G. *The Advanced Theory of Statistics*. Vol. I. London: Charles Griffin and Co., 1945.

RECENT PUBLICATIONS RECEIVED

- BARAHAL, GEORGE D. *Converting a Veterans Guidance Center*. Stanford: Stanford University Press, 1950. 100 pp. \$1.50.
- HAMILTON, KENNETH W. *Counseling the Handicapped in the Rehabilitation Process*. New York: The Ronald Press Co., 1950. 296 pp. \$3.50.
- PORTER, JR., E. H. *An Introduction to Therapeutic Counseling*. Boston: Houghton Mifflin Co., 1950. 223 pp. \$2.75.
- SPEARMAN, C. AND JONES, LL. WYNN. *Human Ability*. London: Macmillan & Co., Ltd., 1950. 198 pp. \$2.50.
- ULETT, GEORGE. *Rorschach Introductory Manual*. St. Louis: Educational Publishers, Inc., 1950. 48 pp. \$3.00.
- Proceedings of the 14th Annual Guidance Conference held at Purdue University, April 4 and 5, 1949*. Studies in Higher Education, LXIX. Lafayette: Division of Educational Reference, Purdue University. 80 pp. \$1.50.

THE CONTRIBUTORS

Marcia Boynton—M.A., George Washington University. Principal Assistant, Head Assistant, Personnel Research, Research Division; Program Reviewer, Personnel Utilization; Examiner, Test Development Unit, U.S. Civil Service Commission. Associate Member, American Psychological Association. Member, D.C. Psychological Association, Society for Personnel Administration

William R. Birge—B.A., Princeton University, 1941. With the U. S. Navy, 1942-1946. Graduate student, Duke University, 1946-1950. Instructor, Rensselaer Polytechnic Institute, 1950—.

Claude E. Buxton—Ph.D., University of Iowa, 1937. Instructor, University of Iowa, 1937-1938. Research Associate, Swarthmore College, 1938-1939. Instructor, Northwestern University, 1939-1942. Assistant Professor, University of Iowa, 1942-1946. Associate Professor, Northwestern University, 1946-1949. Professor, Yale University, 1949—. Author of articles on human and animal learning, methodology, and on the teaching of psychology. Fellow, American Psychological Association. Member, Society of Experimental Psychologists, Sigma Xi, American Association for the Advancement of Science, Midwestern Psychological Association, Eastern Psychological Association.

N. M. Downie—Ph.D., University of Syracuse, 1948. Instructor in Biology, Robert College, Istanbul, Turkey, 1936-1939. Instructor in Education and Graduate Assistant, Evaluation Service Center, Syracuse University, 1946-1948. Assistant Professor of Education, State College of Washington, 1948—.

Frank M. du Mas—M.A., University of Virginia, 1941. Graduate Student, University of Virginia, 1941-1942. War work and military service, 1942-1945. Instructor in Psychology, University of Denver, 1945-1947. Research Assistant, University of Iowa, 1947-1948. Associate Professor of Psychology, Florida State University, 1948—. On contract, Office of Naval Research, under the guidance of the American Council on Education.

Joseph V. Hanna—Ph.D., New York University, 1928. Instructor, Assistant Professor, Associate Professor, New York University, 1926-1949. Director, Veterans Advisement, Vocational Service Center, Y.M.C.A., New York City, 1944-1947. Consultant to Vocational Service Center, at present. Co-author of *The Dissatisfied Worker* and author of articles in professional journals. Fellow, American Psychological Association. Member, N. Y. State Psychological Association, N. Y. Association for Applied Psychology, National Guidance Association. Diplomate, American Board of Examiners in Professional Psychology.

Joseph C. Heston—Ph.D., Ohio State University, 1941. Science

Instructor and Director of Remedial Work, West Jefferson, Ohio, High School, 1932-1939. Assistant in Psychology, Ohio State University, 1939-1941. Instructor in Psychology, 1941-1942; Assistant Professor, 1942-1946, Director of Bureau of Testing and Research, 1944-; Associate Professor, 1946-1950; Professor, 1950- De Pauw University. Research Consultant, Farm Security Administration, 1940-1942. Author of the *Heston Personal Adjustment Inventory* (World Book Co.) and of articles in professional journals. Member, American Psychological Association, American College Personnel Association, American Association of University Professors, Sigma Xi.

J. W. Holley—M.A., University of Southern California, 1947. Counselor, University of Chicago, 1947. Instructor in Psychology, University of Illinois, Chicago Undergraduate Division, 1947-1948. In charge of Admissions Testing, Northwestern University, 1948-1949. Member, Psi Chi, Sigma Xi.

L. J. Lins—Ph.D., University of Wisconsin, 1946. Teacher, rural school, Highland, Wisconsin, 1939-1940. Teacher and Principal, City Elementary, Mineral Point, Wisconsin, 1940-1942. Teacher and Director, Visual Education, Township High School, Amboy, Illinois, 1942-1943. Teacher, Central High School, Madison, Wisconsin, and Research Assistant, Teacher Personnel Research Bureau, University of Wisconsin, 1943-1944. All-University Fellow, School of Education, University of Wisconsin, 1944-1945. Instructor, 1945-1946; Assistant Professor, 1946-1947, Dept. of Education, University of Detroit. Assistant to Executive Director, Bureau of Guidance and Records (Asst. Prof.), 1947-1948; Director, Office of Statistics and Research (Asst. Prof.), 1948-, University of Wisconsin. Author of articles on audio-visual education, teacher education, personnel, and measurement. Member, Phi Delta Kappa, American Statistical Association, Society for the Advancement of Education, American College Personnel Association, Wisconsin Education Association.

Milton M. Mandell—B.A., New York University, 1933. Assistant Director of Examinations, Los Angeles City Civil Service Commission, 1939-1940. Classification Consultant, State of Connecticut, 1940-1941. Regional Personnel Officer, OEM, 1941-1942. Personnel Officer, Office of Program Vice-Chairman, War Production Board, 1942-1943. Chief Analyst, Committee For Congested Areas, 1943-1944. Chief, Administration and Management Testing, U. S. Civil Service Commission, 1944-. Member, American Society of Public Administration, Civil Service Assemblé.

C. Robert Pace—Ph.D., University of Minnesota, 1937. Instructor and Research Associate, General College, University of Minnesota, 1937-1940. Research Associate, Commission on Teacher Education, American Council on Education, 1940-1943. Head, Research Unit and Field Research Section, Bureau of Naval Personnel, 1943-1947. Associate Director, Director, Evaluation Service Center, Syracuse University, 1947-. Author of *They Went to College* (Univ. of Minnesota Press) and co-author of *Evaluation in Teacher Education* (American Council on Education); author of articles on attitude measurement, evaluation, and higher education. Fellow, American

Psychological Association. Member, American Educational Research Association, National Society for the Study of Education, American Association for Public Opinion Research.

Robert G. Smith, Jr.—M.A., University of Florida, 1947. Teaching Assistant and graduate student, University of Illinois, 1947-. Author of work on Member, Phi Kappa Phi, Sigma Xi.

Robert M. W. Travers—Ph.D., Columbia University, 1941. Research Associate, Teachers College, Columbia University, 1938-1941. Instructor in Psychology, Ohio State University, 1941-1943. Personnel Technician, Adjutant General's Office, 1943-1945. Assistant Director, Graduate Record Examination, 1945-1946. Associate Professor of Psychology and of Education, and Chief, Evaluation and Examinations Division of the Bureau of Psychological Services, 1947-. Author of articles on problems of evaluation, statistical methods related to the construction and use of tests, and of a book entitled *Teacher-Made Objective Tests of Achievement*. Associate Member, American Psychological Association. Member, American Educational Research Association.

Maurice E. Troyer—Ph.D., Ohio State University, 1935. Superintendent, Bureau of Township Schools, Princeton, Illinois, 1925-1929. Assistant Professor of Psychology, Bluffton College, 1930-1932. Instructor in charge of Remedial Program, Ohio State University, 1933-1936. Assistant Professor of Education, Syracuse University, 1936-1939. Associate Professor, 1939. Associate in Commission on Teacher Education, American 1940-1943. Director, Bureau of School Services, Professor of Education, Syracuse University, 1943. Director, Evaluation Service Center, Syracuse University, 1945. Member, American Psychological Association, American Association of Applied Psychology, American Educational Research Association, American Association for the Advancement of Science.

Wimburn L. Wallace—Ph.D., University of Michigan, 1949. Assistant, Department of Psychology, 1939-1941, 1949; Assistant Clinician, Psychological Clinic, 1940-1941, University of Michigan. Senior Instructor, Curtiss-Wright Technical Institute, Glendale, California, 1941-1944. Personnel Officer, U. S. Navy, 1944-1946. Chief, V. A. Guidance Center, 1946-1948; Research Associate, Evaluation and Examinations Division of the Bureau of Psychological Services, 1948-1949, University of Michigan. Director of Guidance, University of Massachusetts, 1949-. Member, American Psychological Association, American College Personnel Association, Sigma Xi, Phi Sigma, Phi Delta Kappa.



TABLE OF CONTENTS

VOLUME TEN, NUMBER THREE, PART 2, 1950

<i>The 1950 Convention Program of the American College Personnel Association.....</i>	443
<i>American College Personnel Association, Officers and Committees.....</i>	448
<i>Editors' Foreword.....</i>	450
<i>Developments in Counseling by Faculty Advisers. CARROLL L. MILLER.....</i>	451
<i>Developments in Residence Hall Counseling. MERLE M. OHLSEN.....</i>	455
<i>Developments in Counseling Bureaus and Clinics. ROYAL M. EMBREE.....</i>	465
<i>No Vain Imaginings. THELMA MILLS.....</i>	476
<i>Evaluation and Research in Group Dynamics. KENNETH F. HERROLD.....</i>	492
<i>The Creation of an Effective Faculty Adviser Training Program Through Group Procedures. IRA J. GORDON.....</i>	505
<i>A Genetic Study of Sociality Patterns of College Women. DAVID S. BRODY.....</i>	513
<i>How to Go About the Process of Evaluating Student Personnel Work. WILLIAM M. GILBERT.....</i>	521
<i>Major Limitations in Current Evaluation Studies. RUTH STRANG.....</i>	531

<i>An Inventory of Student Reaction to Student Personnel Services.</i> ROBERT B. KAMM.....	537
<i>The Measurement of Student Conceptions of the Role of a College Advisory System.</i> EDGAR Z. FRIEDENBERG.....	545
<i>The Role of Student Government in the Student Personnel Pro- gram.</i> BROTHUR IGUIS	569
<i>Student Personnel Work and the National Student Association.</i> GORDON KLOPF	577
<i>Contributions of the Student Union to the Total Personnel Pro- gram.</i> DONOVAN D. LANCASTER.....	585
<i>Major Issues and Trends in the Graduate Training of College Personnel Workers.</i> W. W. BLAESSER AND CLIFFORD P. FROELICH.....	588
<i>Employment Outlook for the 1950 Crop of College Graduates.</i> EWAN CLAGUE.....	596
<i>Our Stake in the Occupied Countries.</i> HAROLD E. SNYDER....	601
<i>Plans for the New International Christian University in Japan.</i> MAURICE E. TROYER.....	603

THE 1950 CONVENTION PROGRAM

SUNDAY, MARCH 26

EXECUTIVE COUNCIL MEETING, ACPA

MONDAY, MARCH 27

GENERAL SESSION

PresidingHILDA THRELKELD
Dean of Women, University of Louisville

Symposium: "Counseling Problems and Techniques: Developments for the Future in the Light of an Evaluation of the Present."

"Developments in Counseling by Faculty Advisers"

CARROLL MILLER, Assistant Dean of College of Liberal Arts, Howard University

"Developments in Residence Hall Counseling"

MERLE M. OHLSEN, Associate Professor of Education, Washington State College

"Developments in Counseling Bureaus and Clinics"

ROYAL B. EMBREE, Assistant Director, Counseling Bureau, University of Texas (Read by Gordon Anderson, Director of Counseling Bureau, University of Texas)

LUNCHEON

Presiding.MITCHELL DREESE
Dean of the Summer Sessions and Professor of Educational Psychology, George Washington University

"No Vain Imaginings".THELMA MILLS
Director Student Affairs for Women, University of Missouri, and President ACPA

FIRST BUSINESS MEETING

PresidingTHELMA MILLS
Director Student Affairs for Women, University of Missouri, and President ACPA

Reports:

KATE MUELLER, Chairman Committee on Research

CLIFFORD HOUSTON, Chairman Committee on Standards

GEORGE A. PIERSON, Chairman Committee on Nominations

LYLE W. CROFT, Chairman Committee on Membership

SECTIONAL MEETING

PresidingJACOB H. CUNNINGHAM
Dean of Students, Lynchburg College

"The Role of the Church Related College in Higher Education."

RAYMOND F. McLAIN, President, Transylvania College, Lexington, Kentucky

TUESDAY, MARCH 28

"Council Day"

WEDNESDAY, MARCH 29

SECTIONAL MEETINGS:

"Major Problems of Personnel Administration of Concern to All College Personnel Workers"

1. Presiding DUGALD ARBUCKLE
Director Student Personnel, School of Education, Boston University

Panel Discussion (for those from large universities and colleges)

Panel Members:

MARTIN SNOKE, Assistant to the Dean of Students, University of Minnesota

JOHN I. BERGSTRESSER, Assistant Dean of Students, University of Chicago

DANIEL D. FEDER, Dean of Students, University of Denver

2. Presiding EVERETT B. SACKETT
Dean of Student Administration, University of New Hampshire

Panel Discussion (for those from middle-sized colleges and universities)

Panel Members:

ROBERT KAMM, Dean of Students, Drake University

NATHAN KOHN, Registrar, Washington University

WILLIAM C. CRAIG, Acting Dean of Students, Washington State College

3. Presiding L. R. PALMERTON
Director Student Personnel, South Dakota School of Mines and Technology

Panel Discussion (for those from small liberal arts colleges, church-related colleges, and teachers colleges)

Panel Members:

LAWRENCE RIGGS, Dean of Students, DePauw University

HELEN M. VOORHEES, Director, Appointment Bureau, Mount Holyoke College

LOUISE T. PAINE, Dean, Elmira College

SECOND BUSINESS MEETING

Presiding THELMA MILLS
Director Student Affairs for Women, University of Missouri, and President ACPA

Reports:

PAUL McMINN, Chairman Committee on Publications

RALPH CARLI, Chairman Committee on International Relations

C. H. REUDISILI, Chairman Committee on Proceedings

GENERAL SESSION

Presiding EUGENE L. SHEPARD
Dean of Student Personnel, Stephens College

Main Speech: "Evaluation and Research in Group Dynamics"
 KENNETH F. HERROLD, Assistant Professor of Education,
 Teachers College, Columbia University
 Two Illustrative Studies, reported by:
 IRA J. GORDON, Kansas State College
 DAVID S. BRODY, Montana State College

GENERAL SESSION

Presiding PAUL C. POLMANTIER
 Director University Testing and Counseling Services, Uni-
 versity of Missouri
 Symposium: "Problems of Evaluation in Student Personnel
 Work"
 "How to Go About The Process of Evaluating Student
 Personnel Work"
 WILLIAM M. GILBERT, Acting Director Student Counseling
 Bureau, University of Illinois
 "Major Limitations in Current Evaluation Studies"
 RUTH STRANG, Professor of Education, Teachers College,
 Columbia University
 Two Illustrative Studies, Reported by:
 ROBERT B. KAMM, Dean of Students, Drake University
 EDGAR Z. FRIEDENBERG, Adviser, University of Chicago

SOCIAL HOUR

Hostess ANNA M. HANSON
 Director of Placement, Simmons College

SECTIONAL MEETINGS

(These will be Discussion Groups—no planned speeches—
 attendance at each limited to the first 25 people to apply for
 special admission card at Information Desk. Prerequisite for
 obtaining card is willingness to talk on the topic listed.)

1. Discussion Leader JOHN WITTHAL
 Assistant Professor, Department of Education, Brooklyn
 College
 Topic: "To What Extent Should the Use of Test Results Be
 Limited to Qualified Personnel?"
2. Discussion Leader ROBERT H. SHAFFER
 Assistant Dean of Students, University of Indiana
 Topic: "How Can We as Student Personnel Workers Stimulate
 and Motivate the Student with Higher Ability?"
3. Discussion Leader M. CATHERINE EVANS
 Assistant Director of Counseling, University of Indiana
 Topic: "The Use of Sociometric Techniques in Residence Hall
 Work."
4. Discussion Leader NATHAN KOHN, JR.
 Registrar, University College, Washington University
 Topic: "Are Freshmen Orientation Courses Desirable?"

THURSDAY, MARCH 30

SECTIONAL MEETINGS:

1. Presiding C. W. McCRACKEN
 Dean of Students, Muskingum College
 Symposium: "Student Activities in Relation to College Personnel Work"
 "The Role of Student Government in the Student Personnel Program"
 BROTHER LOUIS, Dean, St. Mary's College, Winona, Minnesota
 "Student Personnel Work and the National Student Association"
 GORDON KLOPF, Chairman, National Advisory Council, N.S.A., University of Wisconsin
 "Contributions of the Student Union to the Total Student Personnel Program"
 DONOVAN D. LANCASTER, President, National Association College Unions, and Director, Moulton Union, Bowdoin College
2. Presiding ROBERT F. MOORE
 Director, Personnel Office, Columbia University
 Panel Discussion: "Reciprocal Contributions of Student Personnel and Industrial Personnel"
 Panel Members:
 DONALD S. BRIDGMAN, Personnel Department, American Telephone & Telegraph Co.
 FORREST H. KIRKPATRICK, Dean of Students, Bethany College
 OTIS C. MCCREERY, Director of Training, Aluminum Company of America

SECTIONAL MEETINGS:

1. Presiding WALTER F. JOHNSON
 Associate Professor, Institute of Counseling, Testing and Guidance, Michigan State College
 Symposium: "Selection and Training of College Personnel Workers"
 Speakers:
 "Problems and Trends in the Selection for Training of College Personnel Workers"
 GEORGE A. KELLY, Director, Psychological Clinic, Ohio State University
 "Major Issues and Trends in the Graduate Training of College Personnel Workers"
 WILLARD W. BLAESSER and CLIFFORD P. FROELICH, United States Office of Education
2. Presiding DONALD J. SHANK
 Vice President, Institute of International Education, New York

Symposium: "Broader Horizons in Personnel Work"

Speakers:

"The Employment Outlook for 1950 College Graduates"

EWAN CLAGUE, Commissioner Labor Statistics, United States Department of Labor

"Aspects of Manpower Mobilization of Significance to College Personnel Workers"

JAMES C. O'BRIEN, Associate Director Manpower, National Security Resources Board

"Our Stake in the Occupied Countries"

HAROLD E. SNYDER, Director, Commission on Occupied Areas, American Council on Education

"Plans for the New International Christian University in Japan"

MAURICE E. TROYER, Vice President in Charge Curriculum and Instruction, Japan International Christian University Foundation

AMERICAN COLLEGE PERSONNEL ASSOCIATION, OFFICERS AND
COMMITTEES

OFFICERS, 1949-50

President: THELMA MILLS, Director, Student Affairs for Women,
University of Missouri
Vice President: E. H. HOPKINS, Vice President, State College of
Washington
Secretary: ROBERT H. SHAFFER, Assistant Dean of Students, Indiana
University
Treasurer: MARCIA EDWARDS, Associate Dean, College of Education,
University of Minnesota

EXECUTIVE COUNCIL, 1949-50

GORDON V. ANDERSON, Director, Bureau of Testing and Counseling,
University of Texas
WILLARD W. BLAESSER, Specialist for Student Personnel Programs,
U. S. Office of Education
EDWARD S. BORDIN, Director, Bureau of Psychological Services,
University of Michigan
DANIEL D. FEDER, Dean of Students, University of Denver
FORREST H. KIRKPATRICK, Dean of Students, Bethany College

OFFICERS, 1950-51

President: THELMA MILLS, Director, Student Affairs for Women,
University of Missouri
Vice President: E. H. HOPKINS, Vice President, State College of
Washington
Secretary: ROBERT H. SHAFFER, Assistant Dean of Students, Indiana
University
Treasurer: MARCIA EDWARDS, Associate Dean, College of Education,
University of Minnesota

EXECUTIVE COUNCIL, 1950-51

GORDON V. ANDERSON, Director, Bureau of Testing and Counseling,
University of Texas
LYLE W. CROFT, Director of Student Personnel Services, University of
Kentucky
CLIFFORD E. ERICKSON, Professor of Education, Michigan State
College
A. BLAIR KNAPP, Vice President, Temple University
DONALD E. SUPER, Professor of Education, Teachers College, Co-
lumbia University

PROGRAM COMMITTEE, 1949-50

CORNELIA D. WILLIAMS, Chairman, Associate Professor and Counse-
lor, General College, University of Minnesota

NORMAN LANGE, Director of Student Personnel, University of Vermont
DUGALD S. ARBUCKLE, Director of Student Personnel, Boston University
JOHN S. BEARD, 5835 Kimbark, Chicago 37, Illinois
LUCILE B. BROWN, Child Education Foundation, New York, N. Y.
RALPH B. BRIDGMAN, President, Merrill Palmer School
HENRY J. CUNNINGHAM, Dean of Students, Lynchburg College
JANICE A. JAMES, Counselor in Occupational Guidance, Stephens College
VICTOR B. JOHNSON, Associate Dean of Men, Clark University
MARGARET RUTH SMITH, Associate Admissions Officer, Wayne University
THOMAS S. RICHARDSON, Director of Student Personnel, Texas Christian University
ALBERT S. THOMPSON, Associate Professor of Education, Teachers College, Columbia University

CONVENTION COMMITTEE CHAIRMEN, 1949-50

JAMES A. MCCLINTOCK, Director of Personnel, Brothers College, Drew University, Local Arrangements
WILLIAM M. WISE, Dean of Student Personnel, University of Florida, Exhibits
HELEN M. VOORHEES, Appointment Bureau, Mt. Holyoke College, Information
MARY D. BIGELOW, Chairman of Advising, Stephens College, Meals
ANNA M. HANSON, Director of Placement, Simmons College, Hospitality
ROBERT H. SHAFFER, Assistant Dean of Students, Indiana University, Publicity
JOHN H. CORNEHLSSEN, JR., Professor of Education, Department of Guidance and Personnel Administration, New York University, Meetings
CLARK I. DAVIS, Dean of Men, Southern Illinois University, Placement.

ACPA COMMITTEE CHAIRMEN, 1949-50

KATE HEVNER MUELLER, Indiana University, Research
CLIFFORD HOUSTON, University of Colorado, Standards
GEORGE A. PIERSON, University of Utah, Nominations
LYLE W. CROFT, University of Kentucky, Membership
PAUL McMINN, University of Oklahoma, Publications
RALPH A. CARLI, Stevens Institute of Technology, International Relations
C. H. REUDISILI, University of Wisconsin, Proceedings
RALPH BRIDGMAN, Merrill Palmer School, Public Recognition
WRAY H. CONGDON, Lehigh University, Local Arrangements

EDITORS' FOREWORD

The twenty-third annual meeting of the American College Personnel Association was held at Atlantic City from March 27 to 30, 1950, in cooperation with the constituent members of the Council of Guidance and Personnel Associations. The convention program was organized to develop the theme, "The Personnel Profession: Achievements and Objectives." Twenty papers were read, eight panel discussions were presented, and two business meetings were held by ACPA members during the four-day period. Eighteen of these papers appear in this publication of the Proceedings. Two papers were not prepared for publication by their authors. The panel discussions held during this convention were not recorded for these proceedings.

On Tuesday, March 28, the members of ACPA participated in the program sponsored by the Council of Guidance and Personnel Associations. At the morning session President Howard R. Beattie made his Annual Report, after which Thelma Mills, ACPA President, and the members of her Committee to Consider Unification made an important proposal to reorganize CGPA into an International Personnel and Guidance Association. At 11:00 a.m. the convention was broken down into many small groups where the proposal was explained further and discussed freely. The convention then reconvened and accepted the recommendation of the Committee on Unification that the reorganization proposal be taken back to the members of the various Associations for their consideration during the coming year and that final action be postponed until the 1951 convention.

At the "Council Day" luncheon meeting, Mr. Laurence A. Appley, President of the American Management Association, discussed the subject, "Greater Utilization of the Educator's Knowledge of Human Potential." In the afternoon, Dr. John E. McGowan, Lecturer in Psychiatry at New York and Columbia Universities, addressed the convention on the topic, "Psychiatry for Counselors." Later Mr. William Line, Professor of Psychology at the University of Toronto, spoke on the subject, "The Scientific Status of Counseling." The papers presented by Mr. Appley and Mr. Line will appear in the *Journal of the National Association of Deans of Women*.

The American College Personnel Association members present at the convention were informed by the Membership Chairman, Mr. Lyle Croft, that our organization is now approaching a total membership of one thousand college personnel workers. With this increase of almost three hundred associates during the past twelve months, we are looking forward to another successful year and to the twenty-fourth annual meeting of the Association which will be held at Chicago March 26 to 29, 1951.

GEORGE A. PIERSON
University of Utah

CATHERINE M. NORTHROP
University of Denver

DEVELOPMENTS IN COUNSELING BY FACULTY ADVISERS

(An Abstract)

CARROLL L. MILLER

Assistant Dean of the College of Liberal Arts, Howard University,
Washington, D. C.

SIGNIFICANT among the recent trends in higher education is a growing recognition of the obligation of the college or university to each student accepted for admission. One result of this development is an increased awareness of the need for "individualization" and the necessity for expanding the facilities for handling the entrant as a person.

The organization of these services may vary from institution to institution, but the aim is basically the same; namely, to assist in the development of the potentialities of the individual within the framework of the philosophy of the school. For the realization of this aim, the college relies in part on its counseling and advisory facilities, which normally include the services of faculty advisers, residence hall counselors, and specialists in counseling and clinical techniques.

The success of any program of counseling in college depends in a large measure upon the effectiveness of faculty advisory¹ services, for the bulk of the counseling problems on a campus are those needing educational guidance, and the faculty adviser is frequently sought by the student when questions relating to academic matters arise.

In order to determine the role played by faculty advisers in the student personnel programs of institutions of higher learning in the United States, a Questionnaire was sent to 115 selected colleges and universities. Replies were received from

¹ The term, faculty adviser, is used here to refer to the general adviser rather than the major field adviser. It is felt that the results of an effective faculty advisory service during the freshman and sophomore years will decrease to a minimum the problems for major field adviser in subsequent years.

90 of these schools.³ The instrument was devised specifically to discover (1) the methods used to select faculty advisers; (2) the services performed by faculty advisers; (3) the methods of orienting and training faculty advisers.

Faculty advisers were available in 86 of the 90 schools reporting. These advisers were selected by the individuals or groups listed below:

<i>Selections made by</i>	<i>Number of Schools</i>
Dean of the College	24
Heads of Departments and Dean of College.	15
Dean of College and Coordinator of Counseling.	8
Heads of Departments.	6
Coordinator of Counseling	6
Dean of Students.	4
Dean of College and Faculty Committee.	3
Heads of Departments and Coordinators of Counseling.	3
Board of Advisers.	2
Dean of Freshmen.	2
Dean of Men.	2
Student Groups.	2
Chairman of General Education Program.	1
Coordinator of Counseling, Dean of Men, and Dean of Women.	1
Dean of College, Dean of Men, Dean of Women, Registrar and Director of Guidance	1
Dean of Students and Heads of Departments.	1
Dean of Students for College, Staff, Chairmen, Dean of College, Dean of Students for University.	1
Dean of the University.	1
Heads of Departments, Dean of the College, and Coordinator of Counseling.	1
President of the College.	1
President, Dean and Faculty Committee.	1
Total.	86

³ Of the 90 institutions from which Questionnaires were received 76 were coeducational; 79 were members of the American Association of Universities; approximately half had faculty members belonging to ACPA. These schools were distributed as follows: New England States 9, Middle Atlantic States 19, Central States 40, Southern States 16, Western States 6.

In selecting advisers four characteristics were taken into account ~~and were reported as follows:~~

Genuine interest in and understanding of students—~~mentioned 73 times.~~

Willingness to take time to advise students without additional compensation—~~mentioned 14 times.~~

Knowledge of course requirements, curricula, and regulations—~~mentioned 8 times.~~

Interest in total educational program—~~mentioned 4 times.~~

The services performed by faculty advisers in 86 schools ranged all of the way from assisting students in selecting courses to helping students gain insight into their personal problems. The activities reported in which advisers engaged are listed below:

<i>Activities</i>	<i>Number Reporting</i>
Assistance in the selection of courses.	86
Assistance in long range academic planning for a career. . .	83
Explanation of academic regulations.	77
Referrals to other agencies.	70
Follow-up of academic progress through periodic reviews of records.	65
Exploration of personal problems.	51
Assistance in securing aids to academic adjustment. . . .	33
Entertainment (social) of advisees.	2
Rating of each advisee on citizenship	1
Assistance in personalizing freshman week.	1

Some form of in-service training for faculty advisers was provided in ²⁴~~53~~ of the institutions reporting.¹ Periodic meetings in which common problems were discussed was the most frequent in-service training method.² Other techniques used were workshops, case conferences, organized summer courses, and faculty adviser's handbooks.

³In the majority of colleges and universities (75) no reductions in teaching load were made to compensate for the time spent as advisers.⁴ Additional compensation was provided faculty advisers by eight institutions;⁵ one institution freed advisers from committee work; and ⁶another institution provided additional compensation and reduced the teaching load.

A few colleges and universities have made definite efforts to improve their faculty advisory services. Among these are Stephens College, where an elaborate adviser's training program is in effect; Ohio State, where advisory services have been centralized; Colgate, where graduate students are used to supplement the services of faculty members; and San Francisco State, where an instructor-advisory plan is now in operation.

While there is greater concern for the welfare of the individual college student today than was true a generation ago, indifference still characterizes the efforts of many faculty advisers. Among the problems yet to be solved are the following: How can faculty advisers be used most effectively? That is, how can their services be made a part of the student personnel program of the institution? What are the personal characteristics of an effective adviser? How important is training in developing an effective faculty adviser? To what extent should faculty advisers attempt to counsel students regarding their various adjustment problems? And, finally, what consideration can and should be made to compensate faculty advisers for their additional responsibilities?

DEVELOPMENTS IN RESIDENCE HALL COUNSELING

MERLE M. OHLSEN

Associate Professor of Education, Washington State College, Pullman, Washington

HAVE you been following the professional literature which has been written on the topic of residence-hall counseling? If you have followed it carefully over the last twenty years, you have found that it has not consumed much of your time. It is true that writers in the field of student personnel work do mention the topic occasionally. They usually agree that the residence-hall program has an important place in the student personnel program.

In preparing this paper it occurred to me that there is one general objective of dormitory counseling. It is to help the student to better understand himself and his relations with people through his day-to-day contacts with interesting and friendly individuals who can work and plan with him. The purpose of this paper is to consider some of the issues involved in achieving this broad objective. Specifically, the following issues will be considered:

1. How are present dormitory counseling services affected by the historical developments in student housing?
2. How does the dormitory staff fit into the general framework of counseling services?
3. What are some of the services which the dormitory counselors can provide?

Let us consider these issues in the order in which they were stated. Stewart¹ reported that the problem of student housing dates back to the very beginning of the great European Universities. This fact in and of itself is not so important, but her account of the gradual shifts in the student's role in house government does have a direct bearing upon student-staff relationships. She traces the change as follows: ". . . in the

¹Helen Q. Stewart. *Some Social Aspects of Residence Halls for College Women*. New York: Professional and Technical Press, 1942, p. 5.

mental flowering and freeing of the Renaissance, residence halls were largely student governed; and that little by little, as learning formalized, authority for the conduction of the life in them was removed from student hands until it rested completely with the college authorities."² If we can accept the statements of philosophy in present-day dormitory staff manuals as an indication of change in practice, it would appear that we are now moving in the direction of more democratic student-staff planning within dormitories.

In any case, we cannot treat the development of the philosophy of student personnel work as it pertains to residence-hall groups as if it were independent of the rest of the student personnel program. Relative to the beginning of student personnel work in this country Cowley³ said that the first college dean seems to have given most of his attention to disciplinary problems. Now if we recall that the early dean often lived in a dormitory as proctor, we see even more clearly why there was a staff-dominated relationship. It is probable that the pattern which was set in these early programs may still plague our dormitory counseling programs today.

It is not likely that the dormitory counselor does his best work if he still holds the "papa or mamma knows best" attitude. We need professional leaders who can work with the students in helping them make plans rather than leaders who devise the plans and attempt to sell them to the students' elected leaders.

The Problem of Dormitory Staff

What has just been said brings to the fore the second issue—that of dormitory staff. It is a problem to find staff members who have the training and the personal security which allows them to work with the students democratically. Orme⁴ said that being a good disciplinarian and "nice woman who loves young people" are no longer adequate qualifications for dormitory heads. Whereas they may have been adequate qualifications

² *Ibid.*, Stewart, p. 93.

³ W. H. Cowley, "Some History and a Venture in Prophecy." *Trends in Student Personnel Work*, E. G. Williamson (Ed.). Minneapolis: Minnesota Press, 1949.

⁴ Rhoda Orme, *Counseling in Residence Halls*, A Report of a Type C Project Doctor of Education Degree, Teachers College, Columbia University, 1948.

for the housemother's position, they are not adequate qualifications for the position of head counselor. Merely liking young people and being kind to them does not qualify the dormitory counselor to provide the kind of services which we shall consider here.

It is true that some of the colleges and universities have met this problem. However, we still do have many housemothers as head counselors. Some schools place teaching faculty in the dormitories as head counselors; others use a combination of teaching staff and undergraduate assistants. Still others staff the living units with graduate counselors. A few employ well-trained, full-time head counselors. The full-time teaching staff member probably is too busy to give the job the time it really takes. Moreover, the job usually demands his attention at the time of day when he prefers to be doing something else.

For the schools which do have the doctoral program, the mature doctoral candidate in student personnel, who has had personnel experience, appears to be the most promising candidate for the head counselor position. First, he has special training. Second, he needs the experience and he is motivated to do a good job. Third, he will be on the job at least three years. His services can be supplemented with upper-class undergraduate students. The young graduate student who is working on a half-time assistantship rounds out the staff nicely. I shall not treat either the problem of the number of staff members needed in a dormitory or the exact qualifications each should have. However, I shall define a given dormitory situation, describe the staff, and treat the problem of services in relation to these factors.

A Specific Dormitory Situation

Now let us think about the specific situation. I shall assume that the hall houses one hundred students. It has a full-time Head Counselor. To assist him in the more specialized services he has a Counseling Assistant who works half time in the residence-hall program and does half-time graduate work. There are also five carefully chosen undergraduate assistants who serve without pay. They act as liaison workers between the students and the paid staff. The Head Counselor has

overall responsibility for the dormitory. Naturally, it is up to him to develop a training program for his own staff. Not only has he the responsibility for training the six staff members described above, but he is also responsible for helping each of the dormitory officers to define and learn how to carry out the duties of his office.

It is obvious that we should know more about the given situation before we attempt to plan a counseling program for it. We should have more information about the students who live there, the kinds of people the individual members of the staff are, and the arrangement of the dormitory itself. But to know that these elements are important suffices for our purposes here.

In passing, something should be said about behavior problems. I believe that we should help students to take responsibility for their own actions. If a student is accused of breaking the social code for the house, his case should come to the attention of the Head Counselor. He, in turn, would help the house officers to collect the facts about the alleged violation. On the basis of the facts, the House Council would make a decision on the case. Should they decide that the case is something which is too difficult for them to handle, the student would be referred to the student-faculty discipline committee. In any case, the house officers should keep detailed notes on the case and the disposal made of it.

Working Relationships

Since we are thinking about the staff, probably I should comment on student-counselor relationships. Even in the residence halls in which students really have had a chance to experience democratic planning, the feeling between the Counselors and the students is different from that in the Counseling Center. The dormitory staff member and the student are personal friends. The dormitory is the home-away-from-home. Hence, we have more of a friend-to-friend counseling relationship rather than a clinical relationship. Here the friendly staff member tries to help the individual students solve problems either individually or in groups. The

staff member not only tries to help individuals, but he tries to set up situations in which students can help each other.

The whole problem of student-counselor relationship also identifies the need for more reflection on the issue of the student's role in the house government. It is my own conviction that democratic planning not only helps to create a better within-house feeling, but it also stimulates greater personal development of the students. If we mean to work with students democratically we must trust them and their judgments. We must be willing to take chances and even to allow them to make mistakes. They must feel that they can settle issues through democratic processes and even go ahead to try a project which the Head Counselor has verbally opposed. This does not mean that the staff leader is not a participating member of the group. He is a member of the living group and as such he has the right to state his arguments in the case. The point is that the staff member should not insist on having his way. Granted, some may feel that too much has been made of this point, but failure to reach an understanding here often seriously affects other staff-student relationships. It is important that there should be established a feeling of mutual trust—an atmosphere in which students and staff can work together democratically in creating and maintaining a living environment with greatest educational, social and cultural values.

Questions of the Teaching Staff

It is also important that the Dormitory Counselors learn to work with the teaching staff. Many questions have been raised by the teaching staff. Suppose we consider just three questions which I heard a staff raise recently:

1. Just what is it that Dormitory Counselors do?
2. Would we be able to notice any difference in our students if these services were discontinued?
3. Is this the best and most economical way of providing these services for students?

We will just have to admit that we do not have the answers to the last two questions now. That means we had better get

busy and evaluate our program. We may need these facts all too soon. The rest of this paper will be devoted to the first question. Just what is it that Dormitory Counselors do?

Duties of the Undergraduate Assistant to the Dormitory Counselor

The undergraduate assistant acts as a liaison between the students and the staff. He makes his contribution by providing the following services:

1. By helping students to become acquainted in the house—both with the students and the staff.
2. By becoming well acquainted with every student in his section—knowing their special interests, abilities, and problems.
3. By referring students for help.
4. By knowing the student resources in the house for special tutorial help.
5. By distributing information which helps all the students keep well informed on both house and college-wide activities and regulations.
6. By helping to promote good house government.
7. By helping to create and maintain a friendly atmosphere. Obviously, this undergraduate assistant would soon lose his opportunity for real leadership in the dormitory if he ever became an inspector for an autocratic Head Counselor.
8. By recognizing morale problems early—since he works with a smaller group of students in the dormitory, he is able to help the head counselor understand sources of difficulty.

The Dormitory Counselor's Services

We have noted some of the things which the undergraduate assistant does. Now what is it that the Head Counselor and the Counseling Assistants do to help students?

1. The Dormitory Counselor should make himself available to students when they need to talk to a friend about personal problems. Those of us who have worked in dormitory programs know that the Head Counselor and Graduate Counseling Assistant can expect to be visited any time of the day or night. The student who knocks at the door during the night probably is too troubled to either sleep or study. He may need no more

than personal attention at the time when things have gone badly. He probably feels the need to talk to a mature friend. However, the trained Dormitory Counselor realizes that the student may need therapy which goes beyond the scope of his job and his competencies.

2. Students want the Dormitory Counselors to help them with their activities. The Dormitory Counselor should do more than merely help students with the activities they now have. He should try to discover the students' interests, then organize small groups to meet individual needs. Some of these small "cell" groups give a student a chance to achieve a measure of security which in turn helps him find and become affiliated with campus-wide activities.

3. Social programs also provide the staff with another chance to help individuals in groups. Such activities as dinners, teas and coffee hours, dances, lectures, musicales, and discussions, all are a part of social education. These experiences can help the student to learn to live in a group and to appreciate some of the cultural values which a college education should provide. On the other hand, it is possible for the staff to promote a social program which the students neither want nor appreciate. Under these conditions little learning takes place.

4. Inasmuch as the dormitory staff member does have a chance to see a student living in a variety of situations, he can provide facts about the student which helps others who also work with the same student. Dormitory staff members often pick up information about the student's family, his personal problems, health, study skills, special learning problems, and study conditions within the house. Some of these facts which the Dormitory Counselor discovers also help such special college committees as the ones on scholarship and discipline.

5. Dormitory workers can become acquainted with the students who need special help. It is important that the Dormitory Counselor not only recognize these students who need special help but that he also *knows the referral agencies* and techniques of referral. All of us would certainly agree that the Dormitory Counselor must be thoroughly acquainted with each agency and its service before he can make an intelligent referral. The referral agency's staff also has a responsi-

bility for working cooperatively with the Dormitory Counselors. Status difference between these two levels of counseling often complicates this task. Since Dormitory Counselors are involved by the mere fact that they live with the student, they must be kept informed about the student's progress and the part they can play in helping him to insure that both of these Counselors are giving the student integrated help.

6. If the dormitory is to become the student's home-away-from-home, then the staff must help to orient him to college. Ideally, the orientation to college should be started in high school. There should also be a college-wide program for the orientation of the new student. Nevertheless, the dormitory orientation program can be a vital factor in the student's adjustment to college and life away from home. The dormitory staff should help the new student become acquainted with other students and the college program.

7. The exit interview is another natural service of the Dormitory Counselor. Inasmuch as students do frequently drop out of school before they adjust to college work this is certainly a needed service in the dormitory. We should accept the student's decision to drop school and to allow him the freedom he needs to talk out his decision. The very fact that we accept his decision to drop out and try to help him plan for the future often causes him to change his plans and stay in school. This is particularly true when his long-term plans do involve college training.

8. Another problem of adjustment to college is the one of quality of scholarship. Actually, there may be as many as four elements in this problem for students: (1) developing good study conditions in the dormitory, (2) helping students to budget their time efficiently, (3) giving assistance in developing good study habits and study methods, and (4) improving reading skills. Of these four elements the dormitory staff can often help with the first three, but they will usually refer the students to the reading clinic for the fourth service.

9. And, finally, there is one other large area of service in which Dormitory Counselors may give help—in educational-vocational planning. It is true that the teaching faculty should do the academic counseling, and that careful vocational

appraisal should be made with the help of a Clinical Counselor. Even so, the students do talk to the Dormitory Counselors about individual courses and fields of study. Hence, the dormitory staff member should have vocational information available to him. He also needs special job information on the fields of study available to students at the college. On the other hand, the dormitory staff should also refer the student to the college's vocational information library. He certainly will want to refer some of the students to a more specialized counseling service for testing and counseling.

Then, there are certain counseling services which Dormitory Counselors can provide. Obviously, not every dormitory staff will be able to provide help in all of these nine areas. The services the staff in a particular residence hall provides must be determined by the quality of the staff and the services provided by the other student personnel agencies. And since the residence hall program is just one part of the whole college program, I decided to conclude this paper with four questions for which answers are still needed:

1. What are the in-service training needs of your Dormitory Counselors?
2. Are we making use of the personnel techniques developed by other agencies and are we adapting these techniques for use in residence hall programs?
3. Is this the best and most economical way of providing the counseling services defined in this paper?
4. Would the teaching staff notice any change in the students if dormitory counseling services were discontinued?

REFERENCES

1. Borreson, B. J. "Student Housing as Personnel Work." *Trends in Student Personnel Work*, E. G. Williamson (Ed.). Minneapolis: Minnesota Press, 1949.
2. Cowley, W. H. "Some History and a Venture in Prophecy." *Trends in Student Personnel Work*, E. G. Williamson (Ed.). Minneapolis: Minnesota Press, 1949.
3. Dammon, A. H. "Residence Halls for Students." *Trends in Student Personnel Work*, E. G. Williamson (Ed.). Minneapolis: Minnesota Press, 1949.
4. Hayes, Harriet. *Planning Residence Halls*. New York: Bureau of Publications, Teachers College, Columbia University, 1932.

5. Lind, Melva. "The College Dormitory as an Emerging Force in Education." *Association of American College Bulletin*, XXXII (1946), 529-538.
6. Lloyd-Jones, Esther McD. and Smith, Margaret R. *A Student Personnel Program for Higher Education*. New York: McGraw-Hill Book Company, 1938. Chapter XII.
7. Orme, Rhoda. "Counseling in Residence Halls" An Unpublished Report of a Type C Project—Doctor of Education Degree, Teachers College, Columbia University, 1948.
8. *Residence Halls for Women Students*. Washington 6, D. C.: National Association of Deans of Women, N.E.A., 1947.
9. Sifford, C. S. "Evaluating a Residence Hall Counseling Program." *School and Society*, XIX (1949), 452-453.
10. Sifford, C. S. "Residence Hall Counseling." *College and University Business*, III (1947).
11. *Survey of Land-Grant Colleges and Universities*. Bulletin 1930, No. 9, Volume I. Washington, D. C.: U. S. Office of Education.
12. Stewart, Helen A. *Some Social Aspects of Residence Halls for College Women*. New York: Professional and Technical Press, 1942.

DEVELOPMENTS IN COUNSELING BUREAUS AND CLINICS

ROYAL B. EMBREE

Assistant Director, Counseling Bureau, University of Texas, (Paper read by Gordon Anderson, Director, Counseling Bureau, University of Texas)

Introduction

For many years it has seemed to the writer that the first need in any speech or article dealing with student personnel work is for a clarification and definition of the very title itself. This notion was reinforced by Dr. Cowley's justifiably choleric variation upon the semantic theme in the Minnesota publication *Trends in Student Personnel Work*.¹ Therefore, the beginning effort in this paper will be aimed at the provision of some basic premises for the consideration of "Developments in Counseling Bureaus and Clinics."

One of the most striking and productive phases of the personnel-guidance mental hygiene movement during the past two decades has been the establishment of a large number of comprehensive agencies, often on college and university campuses, which were designed to provide professional assistance to people through the channels of self-appraisal and counseling. These organizations, whether they arose under the sponsorship of the community or of an educational institution, have made a tremendous contribution to the meeting of individual developmental needs, not only through their direct service to people, but also through their emphasis upon professional training of staff members, scientific methodology and fundamental research. An effort will be made in the following section to trace the origin and growth of centralized psychological agencies in colleges and universities. The important points to consider now are the facts that (1) these agencies developed with a wide variety of titles,

¹ Cowley, W. H. "Jabberwocky Versus Maturity." *Trends in Student Personnel Work*. Minneapolis: University of Minnesota Press, 1949. Pages 342-349.

and (2) these agencies developed in direct response to the evident needs of their potential clientele and not as planned aspects of total institutional student personnel programs.

This paper will be confined to a consideration of counseling agencies which have been developed by colleges and universities. There has been little agreement with respect to the names given to these organizations. An opportunity to study this matter of nomenclature was provided by the excellent directory of counseling agencies recently released by the Ethical Practices Committee of the National Vocational Guidance Association.¹ Fifty agencies sponsored by institutions of higher learning were included in the *Directory*. Of this number, twenty-three, or nearly half, used the term *center* in their listed titles. The next most popular designation was *service*, used by six institutions. Four listed agencies had titles which included the word *bureau*. In three cases, no title was stated. Other descriptive titles and their incidence were as follows: *department*—3, *clinic*—2, *office*—2, *division*—2, *unit*—2, *laboratory*—2, and *institute*—1.

The counseling agencies listed in the *Directory* included many, but by no means all, of the more active and better-known organizations in the colleges and universities of this country. It is clear that the terms used in the title of this paper are among the less popular ones and that preference is tending overwhelmingly toward the use of *center* in the description of these counseling agencies. It seems reasonable to predict that this preference will continue, since *center* has been very widely used in describing facilities for the counseling of veterans which are rapidly being converted into general college counseling organizations.

The *Directory* also provides some interesting information concerning the second major point made above. Only seven of the fifty listed agencies appear to restrict their clientele to the students of their parent institutions. (It is obvious that agencies which do so limit clientele would be less likely than others to list themselves in the *Directory*.) Approximately

¹ Ethical Practices Committee, National Vocational Guidance Association. 1950 *Directory of Vocational Counseling Agencies*. St. Louis, Missouri: Washington University, 1950. 98 p.

60 per cent of listed centers are open to adolescents and adults outside the institution and about 20 per cent are open to outside clients of all ages and levels of schooling. The median listed fee for non-institutional cases falls between \$20 and \$25. Twenty-seven of the fifty agencies indicate that they counsel veterans under contract with the Veterans Administration. It is clear that the majority of these centers have been developed to serve the needs of a clientele extending well beyond the limits of the institutions which sponsor them. This extension of facilities represents an important public service, but, by strict interpretation, it carries the counseling service beyond the logical limits of a *student personnel agency*. On the other hand, however, thirty of these listed centers provide free service to the students of their parent institutions, indicating that they have been developed, at least in part, to meet intramural student needs.

This prevalent dualism in collegiate counseling centers raises an important point. Many of these organizations are actually student-personnel facilities to only a partial degree, and this is especially true of some of the most extensive bureaus, centers and services. Other functions such as clinical work with children, general adult counseling, industrial consultation, examining, test-scoring and educational research may well occupy the greater share of the agency's time and personnel.

It is proposed that the subject of this paper be reworded as *The Central Counseling Facility* for Students in Colleges and Universities, defined as follows:

A central counseling facility is an integral part of a student personnel program which provides an opportunity for *specialized counseling* by *professional workers* with access to the various technical devices which are being developed in the field of counseling.

Such a facility may be part of a very extensive bureau or psychological service center. It may as well be the counseling office of a small liberal-arts or junior college, manned by a single professionally trained clinical counselor. Actually, there may be several *central counseling facilities* inside the same university, each representing a nuclear development within

some subdivision of the total institutional structure. The size of these centers, and the variety of services provided, will cover a broad range and will be conditioned by the characteristics of the institutions which develop them. The crucial point of the concept offered here is that the *central counseling facility* would be recognized by definition as an integral part of the total student personnel program of the institution, and would be considered separately from the other worthy functions often allocated to agencies which render psychological services in colleges and universities. It would seem probable that such a line of thought should tend to eradicate the rather insular characteristics of many counseling centers, thereby improving their integration with other aspects of the institution's total program of services to students.

The Origin and Growth of Central Counseling Facilities in Colleges and Universities

Counseling centers in colleges and universities have tended to develop around the interests and stimulations of certain individuals and, in most cases, have been organized well in advance of the growth of generalized student personnel programs in their parent institutions. The result has been a widespread effort to meet individual needs, institutional and otherwise, by providing the best possible services in the areas of self-appraisal through measurement and/or the counseling of individual clients.

Perhaps the most satisfactory framework for considering the development of these counseling centers has been provided by E. G. Williamson in the first chapter of his book, *Counseling Adolescents*.³ He proposes that the two great emphases upon counseling to date have been (1) counseling as a vocational guidance and (2) counseling as psychotherapy. A tracing back of the factors involved in the development of central counseling facilities in institutions of higher learning will show that they have tapped these two principal sources.

The emphasis upon vocational guidance was apparent in the organizations developed in communities and school

³ Williamson, E. G. *Counseling Adolescents*. New York: McGraw-Hill Book Company, 1950. 548 p.

systems to meet needs in this area. The early period of organization has been effectively described by Reed.⁴ Probably the earliest establishment was the *Vocational Bureau of Boston* in 1909 under the direct influence of Frank Parsons. This type of service was reproduced many times in other school systems, and, shortly after the close of World War I, there were numerous people in colleges and universities who wished to make this vocational-educational service available to students in general. These people were usually employed by departments of psychology or educational psychology and thus it happened that the vocational and educational services in which they believed tended to develop within the confines of these departments.

The emphasis upon personal problems and therapeutic counseling has also exerted a great influence upon the development of central counseling services in colleges and universities. Members of psychology departments, and especially clinical psychologists, were concerned at a very early date with the individual emotional and developmental problems of college students. Their efforts to meet needs in this area were crystallized under departmental sponsorship and often grew into independent central counseling facilities. In a few cases, leadership in personal counseling originated with and was supported by the student health service of a college or university.

A few specialized references may provide body and color to this discussion. In 1934, Williamson reported on the organization of the University Testing Bureau at the University of Minnesota in 1932.⁵ He described how the Bureau was developed to meet the increasingly complicated needs of students and he outlined the philosophy and procedure of the service in clear detail. He reported that 1,932 cases had been handled during the period 1932-1934, and that these individuals represented a reasonably random sample of the university population. This central counseling facility grew out of the interest and stimulation of Donald G. Paterson who brought

⁴ Reed, Anna Y. *Guidance and Personnel Services in Education*. Ithaca, New York: Cornell University Press, 1944. 496 p.

⁵ Williamson, E. G. "Biennial Report of the University Testing Bureau, 1932-1934." P. 343-351, *Report of the President for the Biennium 1932-34*. Minneapolis: University of Minnesota, 1935.

his war-sharpened breadth of thinking to the Minnesota campus.⁶

Another type of development is represented at Ohio State University. Stogdill, who was a member of the Psychology Department, has reported upon the treatment of cases which dated back into the 1920's.⁷ The writer can vouch personally for Dr. Stogdill's excellent work and he recalls having sat in on a case of hypnotherapy handled by Dr. H. H. Goddard, but he remembers that student services were far from integrated since he also worked during 1930 with Dr. Louella Cole in a program designed to assist students in the improvement of reading and study habits. The clinical approach to student problems at Ohio State moved into the era of Rogers and still exists as a distinctly parallel facility to the Occupational Opportunities Service which is more clearly orientated to educational-vocational problems.

McKinney has described the foundation and development of the "College Adjustment Clinic" at the University of Missouri.⁸ This agency was developed about 1938 as an outgrowth of the Student Health Service. It is understood that it exists at present in tandem with a central counseling facility of a clearly educational-vocational nature which was developed to meet the demands of veteran advisement. The same dichotomy of emotional and vocational-educational services may also be found at the Universities of Chicago and Oklahoma, and elsewhere in the country.

A more comprehensive service is described by Bailey, Gilbert and Berg at the University of Illinois.⁹ This central counseling facility was designed from the beginning to utilize the services of clinical counselors and also of trained faculty counselors who were detailed to educational-vocational work with students.

⁶ Williamson, E. G. *Trends in Student Personnel Work*. Minneapolis: University of Minnesota Press, 1949. 417 p.

⁷ Stogdill, E. L. "A Survey of the Case Records of a Student Psychological Consultation Service Over a Ten-Year Period." *Psychological Exchange*, III (1943), 129-133.

⁸ McKinney, Fred. "Four Years of a College Adjustment Clinic. I. Organization of Clinic and Problems of Counselees." *Journal of Consulting Psychology*, IX (1945), 203-212.

⁹ Bailey, H. W., Gilbert, William M. and Berg, Irwin A. "Counseling and the Use of Tests in the Student Personnel Bureau at the University of Illinois." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, VI (1946), 37-60.

In conclusion, it may be pointed out that the development of central counseling facilities in colleges and universities has resulted from an emphasis upon vocational guidance, upon personal counseling, or upon a combination of these two factors. The actual patterns of development in most cases have been highly individualistic—dependent upon the personalities and viewpoints of the principal influencers of growth. The relative newness in the student personnel scene of college counseling services and their close identification with the persons who founded and developed them account for the wide variations which exist today in matters of philosophy, function and policy.

Present Trends in the Development of Central Counseling Facilities in Colleges and Universities

The most striking trend in the development of counseling services in colleges and universities is the rapidity with which these agencies are being activated on the campuses of this country. It was mentioned above that the *1950 Directory of Vocational Counseling Agencies* listed fifty counseling facilities sponsored by institutions of higher learning. In a few moments, the writer was able to think of twenty-five active college counseling services which he knows of personally and which were not included in the *Directory*. Surely, there are many more. If the broad definition given for a central counseling facility be accepted, one could add to the list a large number of strictly intramural but professionally manned offices in smaller colleges and universities. The development of these counseling centers and services represents one of the most active areas of student personnel work.

Probably no one factor has contributed more to the expansion of college counseling services than the Veterans Administration College and University Guidance Program. The implications of this extensive subsidization of counseling facilities for veterans on college campuses were discussed by Dreese at the 1949 meeting of the American College Personnel Association.¹⁰ He reports that there were 415 centers in cooperating institu-

¹⁰ Dreese, Mitchell. "Present Policies and Future Plans of College Guidance Centers Operating under V. A. Contracts—A Survey of the American Council on Education." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, Part II, IX (1949), 558-578.

tions at the peak of the program and that an estimated 1,000,000 cases had been counseled by March 1, 1949. He endeavored to find the attitude of college administrators toward these services and their plans for the centers following the termination of government contracts. There seems to be no doubt that the V. A. guidance program has been a tremendous stimulus to counseling. Furthermore, about four-fifths of 154 institutions intended to continue the centers as part of their college personnel programs even though only half of these schools had maintained a central counseling facility prior to the establishment of the V. A. service.

Another very vital trend is the rapid professionalization of the staffs of counseling services in colleges and universities. The position of clinical counselor has been clearly defined and, in some institutions, is officially established in terms of training standards, personal qualifications and duties. Reference to the above-mentioned *Directory* indicates that very high standards are being maintained by colleges and universities in their selection of directors and professional personnel for central counseling services. Thirty of the fifty listed college centers are led by persons with doctoral degrees and only one director was without some advanced degree. These directors included thirteen people with ABLEPP diploma, twenty professional members of N. V. G. A. and some thirty-four Fellows or Associates of the American Psychological Association. The professional staffs of these centers included approximately 200 counselors, forty clinical psychologists and 100 psychometrists. The *Directory* provided opportunity to indicate how many professional employees were certified (Professional member N. V. G. A., Associate or Fellow of Division 17, A. P. A., diploma of ABLEPP, State certification). Approximately 50 per cent of the counselors, 80 per cent of the psychologists and 17 per cent of the psychometrists were designated as certified personnel.

A very important development is represented by the growing use of central counseling facilities in the training of graduate students who plan to be clinical counselors. Carefully planned and supervised internship and practicum experience in counseling centers have become the crowning factors in counselor

training in a number of institutions. The central counseling facility, regardless of size, can also make a real contribution to on-the-job training of faculty counselors. In some situations, a planned system of rotating faculty members through tours of duty in the central service has vastly improved their training as semi-professional counselors.

There is no need to elaborate upon the trend toward increasing emphasis upon student personnel research in college counseling agencies. They are admirably situated and excellently staffed for this purpose. Already, the college personnel movement owes a mighty debt to certain of the more established counseling services which have produced a large amount of highly significant research in connection with their studies of students and counseling techniques. There is an unlimited future for development in this area, but careful programming, and cooperative planning by counseling services must be achieved, if optimal results are to be obtained.

The rapid expansion of special services in college counseling agencies is another characteristic of present development. Specialized counselors are being provided to assist students in such areas as reading, study, human relationships, preparation for marriage and marital adjustment. This growing tendency toward specialization results in a sort of clinical approach in which several experts share in the analysis and counseling of the individual when this is demanded by the situation. The field of counseling has become so complex and its literature and techniques so extensive that a certain amount of specialization is necessary. However, caution should be exercised in this connection for overspecialization could dangerously threaten the close personal association so important to a satisfactory counseling relationship.

A final tendency, and a very significant one, is the movement toward the improved integration of central counseling services with the other phases of the total student personnel program. There is much to be done in this area, especially in the case of more insular counseling agencies. The task is simpler with smaller, more flexible central counseling facilities. This matter should be carefully considered by the many institutions which are converting their counseling centers for veterans into

student agencies, since there will in such instances be no established interests and policies to obstruct progress toward the integration of all student personnel services.

The Functions of a Central Counseling Facility

This paper will be concluded by the presentation of a schematic system for outlining the functions of central counseling facilities in colleges and universities.

The responsibilities of such a counseling service may be represented effectively by a pyramid with a tri-lateral base. This sort of diagramming appears reasonable, since the central counseling facility can assume a very vital and focal position in the total personnel program of a college or university. The significance of this position is enhanced by the growing tendency to consider counseling as a basic educational process, a viewpoint which has recently been strongly emphasized by Williamson.¹¹

Three functions or services are suggested by the base of this pyramid. The first, and perhaps the most important, is *Side A*, which represents direct, personal assistance to students through the media of self appraisal and/or counseling. There need be little concern regarding this function, for efforts to meet individual needs have been a characteristic aspect of college counseling services since their origin.

Side B of the base is the essential function of training. There are at least three principal areas of training to which the central counseling facility can and should contribute. One is the continuous responsibility for stimulating and up-grading the staff of the center itself through organized programs of on-the-job training. Another is the training of various counselors in the institution (usually faculty or residential) who are contributing to the total job of individual work with students. The third is the task of providing an opportunity for internship, or practicum experience, for graduate students who are specializing in the field of counseling. This need will arise only in the larger colleges and universities, but when it is possible, the integration of counselor-training

¹¹ Williamson, E. G. *Counseling Adolescents*. New York: McGraw-Hill Book Company, 1950. 548 p.

and central counseling activity can make a real contribution to both training and counseling.

Side C of the base is the function of planned assistance to the other agencies in the institution which are engaged in the general task of counseling students. This represents the most neglected side of the figure. Little progress can be made in this direction until the problems of general integration mentioned above have been worked out. However, it is obvious that the central counseling facility, through its access to personnel data and through the insights and experiences of its staff, can render invaluable assistance to other counselors in the institutional personnel program.

It is proposed that the altitude function of this pyramid be considered as deliberately planned and programmed research. The scientific study of students, and of the efficacy of methods used to assist them, will give body or volume to the entire program suggested above. Research may be directed at any or all of the three basal functions outlined: service to students, training, or assistance to the general and non-professional staff of counselors. The absence of this research emphasis reduces the central counseling facility to a plane surface, without body or volume. The applications of research can vastly enrich any of the approaches which are made to serving the three functions of the central counseling facility which have been outlined here.

In conclusion, it may be stated that the maximal value of central counseling facilities can be attained from the filling out of the pyramid suggested in this paper. There is nothing about the representation which needs to be conditioned by the size or number of employees of central facilities. The small central facility, manned by one counselor, can fill out the pyramid as effectively as the great college counseling center. The important facts are that the central counseling facility should contribute to (1) service to students, (2) training, and (3) assistance to extra-center personnel who are counseling students, and that there should be a dominating scientific approach to all that is undertaken in these areas.

Presidential Address
NO VAIN IMAGININGS

THELMA MILLS

Director, Student Affairs for Women, University of Missouri

THERE is a fable about an ancient King, who, troubled by the economic woes of his people, called upon the economists of his kingdom for advice. Confused by their conflicting theories and counsel, he commanded them to prepare a short and simple text on economics for him. After many months they brought him many volumes replete with charts and graphs. In fury, the King banished half of the economists and commanded the other half to produce a text which he could understand. One after another they made reports that went over his head, and one after another they went into exile. Finally, all but one economist was gone. In fear and trembling, this last economist appeared before the King. "Your Majesty," he quavered, "I have reduced this subject of economics to a single sentence. In nine words I will reveal to you all the wisdom to be distilled from all the economists who once practiced in your realm: "THERE IS NO SUCH THING AS A FREE LUNCH!"

As I speak to you today I am much like the *last economist* because, by asking the guests at the head table to join us and *pay* for their own luncheon, I have proved to them that ACPA economics is no less rigorous. In another way I resemble the last economist, because I have set for myself the task of presenting a composite picture of the aims and aspirations of the presidents of ACPA during the past two decades. From the study of these reports came my title, "No Vain Imaginings," for I found that not only were they sound in their thinking, but, also, profound. They did not vainly hope for their plans to be made realities, as you, too, will see in the next minutes of

¹ From an article in *Steelways* by William J. Grey.

presentation. So settle back and prepare to enjoy a family reunion where we again gather together after the wars (personal, professional and actual) to evaluate what *we* have been doing. A reunion always calls for introducing some of the older members to the newer arrivals in the family circle, as well as to the guests, and our ACPA reunion for the fourth time at "Our" Atlantic City palatial home is no exception to the rule. To introduce all of the new family members to the old would be impossible, for our family has grown from a recorded *ninety* in February, 1932, to 894, paid as of March, 1950. May I recognize the 16 who are still active members of the Association:

Fredericka Belknap
Don Bridgman
A. J. Brumbaugh
Frances Camp
M. D. Helser
J. A. Humphreys
Esther Lloyd Jones
Forrest Kirkpatrick

James McClintock
Harriet E. O'Shea
Luther Purdom
Helen Voorhees
Edith Weir
Mary A. Wegner
Lewis Williams
Robert Woellner

We came of age with our 21st annual meeting in 1948, and so the following year our president, C. Gilbert Wrenn, had us analyzing ourselves to see whether in our adult life we were *socially effective* personnel workers. In "The Fault, Dear Brutus," he asked us to discuss with him the psychological problems and temptations of college personnel workers and to think of some of the possible solutions. Now I am sure that our sixteen long-term members must have met the first of his prerequisites to real maturity, "have fun from our associations with people," or they would not be here today, nor members, continuously, of the Association.

Now let us turn our attention to *the* Association, ACPA, and see how it has accomplished the hopes and aspirations of its twelve presidential leaders through the years. I should like for the record to mention them and their schools.

1923-25	May L. Cheney	University of California
1925-27	Margaret Cameron	University of Michigan
1927-30	Francis F. Bradshaw	University of North Carolina
1930-33	J. E. Walters	Purdue University
1933-35	Karl Cowdery	Stanford University
1935-37	Esther Lloyd-Jones	Teachers College, Columbia

1937-39	A. J. Brumbaugh	University of Chicago
1939-41	Helen Voorhees	Mt. Holyoke College
1941-44	F. G. Williamson	University of Minnesota
1944-47	Daniel D. Feder	University of Illinois
1947-49	C. Gilbert Wrenn	University of Minnesota
1949-51	Thelma Mills	University of Missouri

Twelve presidents, and may I call your attention to the fact that five of them have been women, elected by popular vote. This delineation of presidents has been for the record so that the younger members of the Association may have a ready file of reference.

May I review for you, from the reports, what these representatives of yours hoped for and did accomplish. In February, 1923, a group of persons interested in placement met in Chicago and, as a result, organized in 1924 the National Association of Appointment Secretaries with 79 members. From the beginning it was recognized that placement was only one phase of personnel philosophy and practice. The "personnel idea" was spreading in colleges, and the Appointment Secretaries' Organization seemed the logical one to help pioneer in a growing program. Thus, a committee was appointed in 1926 to work with other groups, including the National Vocational Guidance Association, the National Association of Deans of Women, the Department of Superintendents, the Personnel Research Federation, and the National Committee of Bureaus of Occupations, in the planning of joint meetings. A community of interests rather than any thought of merging into an over-all organization brought these early leaders together.

In 1929, the name was changed from *National Association of Appointment Secretaries* to *National Placement and Personnel Officers*. In 1930, in Atlantic City, a new constitution was proposed and the following year in Detroit the name was changed to *American College Personnel Association*. The new constitution was adopted and sectional divisions were set up in Educational Counseling, General Placement, Personal Counseling, Records and Research, and Teacher Placement.

The 1932 annual meeting was devoted to a "Study of Personnel Activities in Members of the Association." Here I must interpolate that "institutions" were the first members,

hence a study of personnel activities *in members* was in no way a "Dies Committee" hunt nor an F.B.I. investigation. The declared purpose of the Association was to increase the number of *departments of personnel* in Colleges and Universities by offering free advisement with ACPA officers. Ninety-five Colleges and Universities were members of the Association and seventy-one of them returned the data which were used for tabulation. May I quote from the paragraph on trends: "of the 15 college personnel departments expressing a trend regarding *administration of work* of the department, *seven* indicate greater centralization of personnel activities; 12 expressed a trend toward better guidance; 11 reported a trend toward more general employment work; 6 departments expressed a trend toward more and better teacher placement."

Three items were of particular interest in the history of the Association during 1932: there was affiliation with the National Association for the Advancement of Science; the Annual Report was published as a separate publication for the first time; and the Association appointed, at the request of the U. S. Civil Service Commission, a committee to make a study of opportunities for women in government, with Mrs. Chase Going Woodhouse as chairman of the committee

By 1933, we had found that prosperity had permanently disappeared around the corner. Presiding at the tenth annual meeting, Jack Walters described the Minneapolis conference as one of quality rather than quantity, with comparatively few attending because of depression and reduced budgets for traveling expenses. It was a year devoted to the preparation of a clearer statement of personnel principles and functions; to the establishment of higher standards of professional work; and to the search for a practical method of judging the effectiveness of college personnel services. The trend toward effective coordination of associations and agencies interested in guidance and personnel continued. Under the inspiring leadership of Dr. Harry Kitson a Coordinating Committee met with Dr. Keppel of the Carnegie Foundation to seek for ways of unifying the ten Associations "through headquarters, cooperative planning of programs of research, yearly activities and conventions, and joint publications."

It is interesting to note that the need for a permanent secretary was discussed at the 1933 meeting. The present need is even more urgent. This is one of our *vain* imaginings because of budget. Until we raise our dues, or increase membership far beyond the now "nearly one thousand," the acquiring of a permanent secretary will remain in the planning stage.

In 1934 Karl Cowdery stated that the purpose of the year's work was "to approve more cooperative action with guidance and personnel groups." The 84 individual members and 18 institutional members voted approval of this purpose and planned to join the *American Council of Guidance and Personnel Associations* with the following member Associations, four of which still remain members:

- *American College Personnel Association
- Institute of Women's Professional Relations
- *National Association of Deans of Women
- *National Vocational Guidance Association
- National Federation of Bureau of Occupations
- Personnel Research Federation
- Southern Women's Educational Alliance
- Teachers College Personnel Association
- American Association of Collegiate Registrars (Affiliated)
- *National Federation of Business and Professional Women's Clubs (Affiliated)

Research was the dominant theme of the 1934 conference. The papers were definitely slanted toward "the personnel point of view," and, more particularly, to "individualized problems of students."

Dr. Grayson Kefauver, of Stanford University, keynoted the 1935 convention with his address on "Developments In Educational Institutions." The contrast between the mechanistic and individualized philosophies of education was sharply drawn. He made it clear that personnel policies should be formulated in terms of the latter philosophy.

As an Association, this was a year for action. Seven thousand names of college staff members throughout the country, who had responsibility for personnel functions, were contacted to further professional solidarity in the personnel field at the

* Current members of the Council and Guidance Personnel Association.

college level; a formal offer was made of the services of the Association to the federal government "in making and executing plans for the services of youths between 18 and 26 years of age."

The theme of the following year (Problems of Personal Adjustments in Moral, Religious, and Social Relations) reflected the same point of view. J. Hellis Miller, then Associate Commissioner of Education in New York State, raised a fundamental question. "Is personnel work an adjunct to, or is it education itself?" The question was clearly answered. Personnel services are not superimposed upon the educational process; they are an integral part of it.

Vice-President Hopkins gave us the follow-up twelve years later, in 1948, in his paper on "The Essentials of a Student Personnel Program." The whole-hearted response to an individualized philosophy of education, accepting the theory first and then putting it into practice, means that it must be written into the educational philosophy of each institution and considered to be the *means* of education, not adjunct to education.

As an Association, we voted to continue as a member of the ACGPA and request the Council to continue its three committees, (Research, Publications, and Coordination). Dr. J. E. Walter proposed "an investigation into what personnel services are being rendered at present in different colleges and universities." He urged that a committee of three be appointed to initiate research projects such as (1) the advisability of formulating a statement of types of preparation offered for the training of personnel workers, (2) the preparation needed for college personnel work. (Corrine LaBarre made such a report in Columbus, Ohio in 1947.)

Another action worth commenting upon dealt directly with us—the placement needs of our own membership. It was proposed that the Chicago Collegiate Bureau be used as a clearing house for filling personnel positions and placing personnel workers. We are still working on such a program, as you will hear on Wednesday.

The personal element enters into the next step in our history for it concerns my own first attendance at an annual meeting. New Orleans was the place, 1937 the year. Esther Lloyd-Jones

talked on "What is this thing called Personnel Work?" She defined the more immediate needs of the association as:

1. a continued effort to clarify the nature and scope of our professional field,
2. fundamental modification of the constitution to conform to our changing conception of the nature of the personnel program in higher education,
3. and continued, careful, patient but aggressive attempts to cooperate within the A.C.G.P.A. with the guidance and personnel groups in the Council. (Akron).

This was also the year that W. H. Cowley gave us "A Preface to the Principles of Student Counseling," stating three fundamental characteristics of counseling:

counseling as the personalization of education,
counseling as the integration of education,
counseling as the coordination of student personnel services.

He defined counseling broadly, "seeing the student and working with him as a whole person." No vain imaginings, for again we read this as a follow-up on the thinking of ACPA members expressed two or three years earlier.

Our 16th annual meeting was held in Cleveland and with Dr. A. J. Brumbaugh speaking on "Personnel Services in the Light of Current Trends in Higher Education." After presenting to us the "unitary nature" of the early American college, with the basic curriculum, he showed that as time advanced the program of colleges became more diversified, both as to scope and content, and that "fan like, higher education extended wider and wider in more divergent directions." By the 19th century we had denominational colleges, women's colleges, land grant colleges and specialized types like art, business, and normal schools, an increase from the 10 unitary colleges before the Revolution to the 2000 institutions of higher education today. Now, after the first of the 20th century this elective system is indicted in many quarters on the ground that it has led to early specialization at the expense of a broad liberal education. The assumptions of that period point "to a unified and generalized educational experience in direct contrast to the specialization that has prevailed." Just what shall be the nature of *General Education* is still a matter of opinion and experimentation.

A new movement in quite another direction, from that of the return to liberal arts, has developed. The leaders of this movement believe that the essential unity of general education is not achieved through curriculum, but through the educational experience of the individual. Thus, the interests and aptitudes of individual students must constitute the focus of this education in which they acquire a self-discipline, integrates learning with experience, functions creatively in the society in which he lives. These two trends both attempt to achieve an essential unity in college education, one by way of intellectual disciplines, the other by way of individualized educative experience.

The personnel services provided in any college must be based upon the purposes of the college and the needs of the students. Some services will be the same in all institutions regardless of individual differences because some of the student problems and difficulties will be the same, such as: selection of a college, variations in student interest and abilities, choice of vocation, the social development of the students, the health of the student, financial aid. The effective functioning of the intellect depends upon many collateral factors, as well as the free and disciplined intellect.

Functionally, as an Association, the year 1939 was memorable in our history. The CHARTER for the ACPA became a published reality. This charter was drawn by "the Commission on Reorganization of the ACPA" appointed in 1937 and composed of Esther Lloyd-Jones, Karl Onthank, and C. Gilbert Wrenn. Basic to the preparation of the charter was the viewpoint reflected in *The Student Personnel Point of View*, a brochure published by the American Council on Education.

This was the year that a committee with Edith Weir as chairman was appointed to write the history of our Association. At the preceding convention which celebrated the 15th Anniversary of ACPA it was found that only a fraction of the members knew the early thought and effort which brought about our organization. Thus, it seemed time to review our past and to secure, from the early members, the information which only they possessed.

The history was to be a compact record covering the various periods of growth from problems of teacher placement to the broader personnel phase, and the effort to develop programs covering various fields of endeavor with the resulting changes

in names. Mrs. Cheney's forty years of placement work had given her invaluable knowledge of early personnel development not to be obtained from any other source, and it was felt that this should be recorded while she was still alive. Mrs. Cheney was given a life membership with full privileges in the Association.

The need for regional groupings was discussed for the first time. May I quote: "to organize regional meetings in many sections implies a great deal of preliminary missionary work on the part of the membership committee." No action was taken. The Committee on Relations with Faculty Advisors reported that the Committee had not yet advanced to making any recommendations concerning an invitation to faculty advisors to become members. The recognition of needs at one convention and their implementation at another have characterized the pioneering of our organization from the beginning. An even more appropriate illustration is found in Daniel Feder's suggestion at the 1940 meeting that we foster the establishment of a journal to print research in educational personnel and other closely related fields. This did not become a reality until 1944.

The St. Louis Convention was held under the leadership of Helen Voorhees. A membership of 239 was recorded. At this meeting a panel prognosticated on "The Future of Student Personnel Work" with the major issues summarized under (1) the functional curriculum and student personnel work, (2) the teacher and personnel work, (3) and the need for a strong national organization. This seemed to be such a forward looking program that I wish to bring the summary, via the 1950 proceedings (Vol. XVII, p. 19), to you in full.

The panel discussants included: A. J. Brumbaugh, C. F. Malmberg, H. W. Bailey, H. D. Bragdon, D. Stratton, and H. H. Moreland

The major points of issue are summarized under the following headings: The functional curriculum and student personnel work; The teacher and student personnel work; The need for a strong national organization.

The functional curriculum and student personnel work: The present need for student personnel work in our colleges and universities arises largely from a curriculum centered in subject matter rather than in student needs. Furthermore, this cur-

riculum is taught by instructors who are narrowly trained in subject matter areas. The functional curriculum, if and when we adopt it, will probably preclude the necessity for having personnel officers, at least of the same type as at present. This point of view is generally held by most personnel workers. However, curriculum changes never occur with lightning-like rapidity. One discussant who had made a careful and exhaustive study of the history of higher education, maintained that the functional curriculum will not dominate higher education for several centuries. In the meantime, personnel workers have much to do. Others hold that a real possibility exists of a radical change in higher education. If institutions of higher learning do not change from their traditional ways, mounting economic and social pressures will force changes.

The teacher and personnel work: Can teachers be trained in the personnel point of view so as to take over a large number of functions now administered by personnel officers? One point of view maintains that college teachers cannot and will not be trained in personnel methods and viewpoints because of the nature of graduate training and the traditions of research and scholarship. By and large, college faculties are recruited from the graduate schools. Graduate training is oriented toward research, not toward students or teaching. Furthermore, academic rewards are not won by the Great Teacher, but by the Great Scholar.

The opposite point of view, held by a large number of people, is that all teachers should be trained personnel workers. With such additional training, teachers would do a better job of teaching and students a better job of learning. While this is acceptable in regard to secondary teaching, the adherents of the first point of view hold no hope that this can be accomplished at the college level. They maintain that student personnel specialists would still be necessary even with a functional curriculum and with the student point of view. They will grant that the faculty may play a role in the instructional type of personnel work; e.g., remedial reading, how-to-study, etc.

In a few places, faculty members and graduate students who expect to teach are taking courses in methods of teaching, personal counseling, etc. Summer workshops, such as are offered at several centers, are organized to give personnel and teaching experience to college teachers. These innovations are unique, however.

The need for a strong national organization: One viewpoint maintains that college personnel work will always be a sideshow of education unless we have a strong national organization which unifies all the branches of student personnel work. The opposition states its point this way: Strong national organizations have a point and push it. Student personnel work is not yet ready for such a vigorous program. We are still experi-

menting. Let us not hamper experimentation by adopting dogmatic attitudes. This is true not only of personnel work but of higher education as well.

The Council of Guidance and Personnel Associations is one attempt at unifying the field and providing a strong national organization. Some hold that it is not broad enough, that such personnel groups as the registrars associations, the health officers, the union managers, etc., should be represented. Others state that all college personnel organizations should unite into one association, divorcing themselves from organizations which are made up predominantly of secondary school people.

If representatives of college personnel services want to form a unified organization, no blueprint is available. They will be obliged to work out their plans in conference, in regional meetings, and in group discussions. First, however, they must study the problem in terms of the needs of personnel work."

It was at this conference that we also had the first emphasis on group dynamics presented. Dr. Ruth Strang reported upon her research in the field of group work and techniques.

Work with groups often constitutes a more successful way than counseling of attaining empathy with individuals, of encouraging them to express their emotional problems, of providing constructive outlets for their impulses and of relieving their tensions and anxiety. . . . Economy is a factor in the development of group work. In counseling, needs of individuals for certain group activities are discovered. Group activities serve as avenues of adjustment, thus they have both diagnostic and therapeutic values.

Atlantic City, February 18-22, 1941, the 18th Annual meeting! Membership, 256. The Association gave serious attention to new membership requirements, "professionally trained persons and other interested, experienced and competent workers." The membership approved of "dignified, slow expansion and growth."

The highlight of this meeting was the presidential address. President Voorhees spoke to us on "The Responsibilities of the Heritage of Personnel Work."

We hear much these days, of the advances which have been made in *personnel methods*, but for the time being, I should like to look back to the past, to the beginning of personnel work. Our predecessors had a rich background in an allied field of education; an experience which had given them a firm con-

viction and belief in the eternal verities. And they had marked success in carrying out their educational aims and purposes. They transferred their sense of values from the pulpit to the field of education. They came to their task possessed of the wisdom which comes only from a wide knowledge of human nature and its frailties; teaching the virtues which are necessary for living and for satisfaction and achievement.

The purpose of this presidential address was to present some aspects of our work which are sometimes forgotten, the spiritual values of our profession.

Great characters, not just great scholars, were produced. Men devoted to service, with initiative, self-reliance and democratic ideas.

Have our methods tended to emphasize personality rather than the necessity inherent in each of us of becoming a person in one's own right?

I am eager that one of our openly avowed objectives shall be to give the young people in our care some philosophy of life which will make it possible for them to get their bearing, no matter what happens.

She was successful both in her presentation and purpose.

In February, 1942, E. G. Williamson faced an especially difficult period of administration. Despite the war, our President led us to think of that future, which lies beyond the present, in personnel work. He showed us that anything which leads to more effective conservation or utilization of youth's potentialities actually does contribute to society's welfare, as well as to the winning of the war. Hence, his address on "The Future Develops Out of the Past" was a highlight.

Whatever our professional and personal behavior is as personnel workers, one thing is quite clear. Unless we have had the benefit of professional training and experience which prove to be effective in our handling of post-war problems, then we may expect that society, including college students themselves, will push us aside and find other types of personnel workers or other types of educational workers to handle this type of social revolution. The pressure for a solution to these problems, the greater articulateness of students and parents and the competition for public favor and support from members of social and government agencies, will force college administrators to deal effectively with this anticipated situation. If we cannot do the job, then others will be found to do it.

I believe that we are adequately prepared for the task and

that we will make an effective contribution to the conservation of useful idealism, realism and social and personal values.

I believe that our contribution will be such as will strengthen our place in higher education and will increasingly attract able graduate students to secure the necessary professional and personal training to make college personnel work a significant part of higher education, competing successfully with other social welfare professions for the best talents in each student generation.

The numbers that have been entering our profession have borne out his faith in the personnel program.

From 1943 to 1946 the only record of the Association is that of the minutes kept by the secretaries and the record of the New York meeting in 1943 when only a limited group met under the leadership of CGPA. Our officers and members were aiding in the promulgation of war programs in every branch of the service; those members left on the campus were doubling as personnel officers and campus recreational leaders for the military services on our campus. A record of these activities would be almost a complete history of the war activities. At a meeting of the Executive Council held in Chicago in December, 1945, a *Personnel-o-Gram* was born, with Fred McKinney named as the first editor. The Council attempted to keep the membership informed through the media of the EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, published by an ACPA member, G. Frederick Kuder. The following ACE brochures, for which ACPA members had been chiefly responsible, were purchased, and distributed to the membership:

"Counseling and Postwar Educational Opportunities."

"Student Personnel Work in the Postwar College."

Active participation in the work of CGPA was continued with special attention given to regional conferences. "Judicious publicity" was carried on by sending a letter to some 1200 college presidents concerning the Association and enclosing a paper written by Dr. John Darley on "Counseling and Colleges in Post-war Education."

By our first postwar annual meeting, held in 1947 at Columbus, Ohio (moved from Chicago, by consent, when the Stevens Hotel would not promise to accommodate *all* our members without discrimination), our Annual Reports were

resumed and issued as a supplement to EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT. During this meeting we were definitely interested in post-war personnel services. In his presidential address, "When Colleges Bulge," Dr. Feder awakened us to the imperative need of making an immediate and reasonably adequate adjustment to things as they are and not as they were in the nostalgic *good old days*.

The problems discussed were common to all of our campuses: (1) the changed nature of campus population, (2) the generally changing motivation and orientation of all college students, (3) the need for high caliber professional services in vocational, educational and personal counseling of all students, (4) special problems, caused by previous military treatment of situations similar to those in classrooms, (5) ways in which the integrated personnel service program may serve both faculty and student body in more effectively meeting student needs. New services were being offered to students in their quest for maturity.

President Feder called our attention to the fact that the field of student personnel work has suffered from an ill of its own making, the tendency to divorce its findings and activities from those of the classroom. As a matter of routine, he insisted we must transmit to the instructional staff those findings regarding student reactions and needs which will assist the faculty in the infusion of the realities, meaning, and purpose of contemporary life in the classroom.

In Chicago, in April, 1948, C. Gilbert Wrenn, our "Chief" spoke to us on the "Greatest Tragedy in College Personnel Work." It is worth our while to review these tragedies briefly for they point the way, just as the meeting, in 1940, on "The Future of Personnel Practices," gave impetus to developments of the early forties. Guidance, as a term, was buried, and personnel, with its appropriate adjective, was nurtured, so that we might speak in common terms with school and non-school agencies about our concepts, as written into the Charter of the Association. Counseling was relegated to its appropriate position as *one* of a number of personnel functions and not *the* entire personnel program.

Outstanding among these developments are the increased participation and consequent demand for professionally

equipped personnel workers; increased facilities for personnel research; and not least important, an increased humility on the part of all of us.

He was not less concerned to point out pressing problems needing to be resolved:

1. The lack of commonly accepted *standards of performance* and professional preparation.
2. Students and faculty, who have the most to gain from student personnel work, have the least to say about its development and emphasis.
3. Poor coordination of a student personnel program is frequently the result of an incompletely formulated line and staff organization.
4. A student personnel program on a campus tends to be isolated from four important influences in the life of the student, (a) home, (b) secondary schools, (c) college classroom, (d) spiritual resources of the campus.

As a final imagining I wish you to think briefly with me about human relationships, a field in which some of us may be devoting more of our time than to the more technical areas. Certainly this is a field in which we can never become complacent with our achievements. The nation's colleges and universities, today, are placing more emphasis on producing well-rounded citizens. How would you answer the provocative question raised by the Pennsylvania Association of Deans of Women, "Do you improve human relationships through your guidance services?" Each of us must answer the question!

When enough people can answer the question in the affirmative we shall indeed have arrived at the place in the personnel profession where we do not have to rely on vain imaginings. One does not need to sentimentalize the point. There is increasing evidence of a deepening unity among individuals and groups devoting themselves to the improvement of human relationships through personnel services.

On a tablet in front of the Old South Meeting House, in Boston, are words that describe our Revolutionary forefathers as "worthy to raise issues." They knew which things were important and which were unimportant; and a person must be mature to raise issues. Most of the small frictions in life, human misunderstandings, that destroy mutual confidence

come from raising issues that are not worth raising, and most of the social inertias and timidities that keep our world from moving toward its ideals express a reluctance to raise issues that should be raised. One of our great responsibilities is to bring more reasonableness into the human scene, to bring it to ourselves and to others. To carry our share of this responsibility we need to see in whole, instead of in part, to be ready to act responsibly where responsibility is called for, to forget ego and to seek wise understanding of others. "Where there is no vision the people perish," and our part is to seek to make mature individuals of ourselves and others, that we may bring about a community where human beings may realize their visions. On a recent drive in my state I passed through three neighboring communities, Vista, Fairplay and Humansville, which have given me a new philosophy for daily living. "Where the vista is right, there will be fairplay in humansville," and you and I must make it come to pass.

EVALUATION AND RESEARCH IN GROUP DYNAMICS

KENNETH F. HERROLD

Assistant Professor of Education, Teachers College, Columbia University

UNDERSTANDING and evaluation of group dynamics must be in terms of the nature of the times in which men live. Industrialization has led to the wholesale, and sometimes indiscriminate, application of the scientific method to the material universe. The phenomenal and, at times, ghastly social and technological changes of this century have led to collective hysteria in one form or another in all parts of the world. The future of social science and, more important, the fate of mankind depend upon whether or not the populations of the world can adjust their living to the atomic era and live and work together intimately and creatively.

The search for a science of human relations is not new. Some have denied that it ever would be possible to apply scientific methodology to the processes of human relations and at the same time to preserve individual freedom and a democratic society. These are legitimate challenges. Others believe that we must develop new forms of social discipline for interpersonal relations. There has always been appropriate skepticism of such suggested social innovations and inventions, and there has always been some inappropriate skepticism concerning the contributions of social scientists like those engaged in group dynamics research.

Group dynamics is a term usually associated with certain concepts and procedures of research and study identified with Kurt Lewin and the Research Center for Group Dynamics first established at the Massachusetts Institute of Technology and later moved to the University of Michigan. However, group dynamics has aroused the interest of other social scientists who have never worked intimately or directly with the Lewinian group. These social science explorers are contributing valuable knowledge to the understanding of how groups

behave. Lewin, Lippitt, White and their associates have represented a strong team. Their work, such as the Iowa studies of the "Social Climate of Groups,"¹ has stimulated thought and controversy which has forced many others to consider basic problems of group and intergroup behavior before they might otherwise have done so. It is necessary and appropriate to indicate that responsible research and training in group dynamics is now being carried on at Harvard, Minnesota, London, California, New York University, Columbia, Northwestern, and many other institutions of advanced study. There is no question of the status of the staff of the Research Center at the University of Michigan. However, to associate the development of group dynamics as a respectable field exclusively with this Research Center is to limit the progress of knowledge and inquiry.

The turbulence set up by the group dynamics enthusiasts has not been universally supported nor accepted. Dean Robert B. Browne recently said of group dynamics:

We are told that we have here something new and basic. One would like to remain receptive to what is new and basic without prejudice. We are told that here is something scientific, from Bethel laboratories and the M.I.T. and Michigan Research Centers. We all have a great respect for scientific inquiry and a staunch faith in its usefulness. We are told that here is something awfully democratic, and that seems to be all to the good. Furthermore, we are assured it's for leaders, which ought to guarantee crowded classrooms where leadership training is offered. But just what is this new, basic, scientific, democratic leadership training furor, and what is there about it that is as yet either new or scientific or democratic or dynamic or even useful?²

Other words of criticism and challenge have been leveled at the proponents of this approach to an understanding of group relations. The development of studies of group dynamics has been accompanied by considerable misunderstanding, misinformation, and erroneous interpretation. The term "group dynamics," therefore, signifies great challenge and

¹ Lippitt, R. and White, R. "The "Social Climate" of Children's Groups." *Child Behavior and Development* (R. Barker, J. Kounin, and B. Wright, Ed.). New York: McGraw-Hill, 1943.

² Browne, Dean Robert B. From an address delivered at the annual convention of the National University Extension Association held in Edgewater Park, Mississippi, May 4, 1949.

hope to some, and to others it represents but a transient panacea.

It has been said that the social unit approach to understanding of human behavior denies the uniqueness of the individual.¹ Need it be dogmatically an "either or" relationship? Can we ever begin to meet student personnel needs on a strictly individual basis? Those of us who are daily confronted with impregnable schedules of individual appointments know how difficult it is to achieve even a satisfactory quality in our counseling relationships. Professional competence in the use of groups and in the analysis of group dynamics can be achieved without detracting from the "student personnel point of view" and without attenuating the warm and friendly relations with students. In fact, the relations of personnel administrators with students, especially students in groups, may become even more respectable the better we understand the behavior of people in groups.

Misunderstanding of the objectives and procedures of group dynamics is, in part, due to the rapid growth of the field and the customary lack of adequate communication which accompanies social innovations. It is also due to the inability or the lack of opportunity adequately to define the nature of group dynamics. This is regrettable. The purpose of this brief paper is to attempt to present: (1) one definition of group dynamics research and application, (2) a citation of certain problems in its developing research and evaluation studies, and (3) a prediction of some of the possible applications of such training, research and evaluation in college personnel administration.

It would seem wise first to understand what those interested and working in the area of group dynamics are trying to do before we examine their research and certainly before an attempt is made to evaluate their activities.

With what is group dynamics concerned?

First, group dynamics is concerned with an understanding of the group related factors, forces, and determinants which

¹ Snygg, Donald and Combs, Arthur W. *Individual Behavior*. New York: Harper Brothers, 1949. P. 183.

influence individual behavior in groups and the course of social change. Lewin described this goal in his statement of the "Frontiers in Group Dynamics"⁴ In general, a group is a social organism of describable structure and function. In most instances the members of a group maintain a face-to-face relationship. Group dynamics is also concerned with the achievement of an understanding of groups as groups and the fundamental laws which govern the behavior of groups. Obviously such studies must rely upon the theoretical and applied experience of all the social sciences, but especially upon social psychology, individual clinical psychology and cultural anthropology.

Second, group dynamics is concerned with improving the application of already established knowledge and skills of human behavior to the critical social problems of our times. This objective was made explicit in Lewin's second statement on the "Frontiers of Group Dynamics."⁵ Impatience with the manner in which social action has developed led many to challenge applied social science in recent years. Sellitz and Cook expressed this concern in their inquiry "Can Research in Social Sciences Be Both Socially Useful and Scientifically Meaningful?"⁶ It has been stated that it requires fifty years for society to adopt a new idea or practice. Much of the knowledge of theory and practice now being utilized by those concerned with the study of group behavior was developed and established many years ago. Some of the psychologists and social scientists who are today most critical of the "group dynamics" movement assisted in the early discovery and definition of social phenomena which form the theoretical base upon which the group dynamics movement is established, and these critics continue to reiterate the same or similar concepts with respect to social action, with no awareness of their commonality with group dynamics. It is often quite a different matter to apply social science than it is to write

⁴Lewin, K. "Frontiers in Group Dynamics: Concept, Method and Reality in Social Science; Social Equilibria and Social Change." *Human Relations*, I (1947), 5-41.

⁵*Ibid.*, I (1947), 143-153.

⁶Sellitz, Clair and Cook, Stuart, W., "Can Research in Social Science Be Both Socially Useful and Scientifically Meaningful?" *American Sociological Review*, XIII (1948), 454-459.

about it or to discover its laws in the protection of controlled laboratory conditions. In fact, it is always difficult to apply what we believe, to reduce the lag between our knowledge of social science and our application of that knowledge to problems of human relations and social advance.

The *third* aspiration of those concerned with group dynamics is to enlist other social scientists from a variety of disciplines in the further study of group development, interpersonal relations within groups, relations between groups, and the basic laws of human relations. This requires an unusual type of intellectual maturity and material; it requires a complete integration of the basic theory and practice of many disciplines of social science.

The present departmentalization of subject matter and professional training has handicapped this type of integrated study and practice. Furthermore, the study of group dynamics must be experience-centered and many of our institutions are not yet ready to utilize the community as a laboratory for advance study and skill development. Time and space will not even permit their cooperation in integrated study within the institution. Communities uninitiated to such working relationships with university or academic men must also be cultivated to utilize such resources. One may speculate, however, whether the community is more ready to have the academic man step outside his ivory tower than is the academic man ready to leave the protection of his isolation. This is, of course, a criticism every generation makes of the scholar and the researcher. Students and professionals in advanced studies in many fields are demonstrating the importance of an understanding of group dynamics and the application of group procedures on their work. A few examples may be cited to document this assertion. Dr. Max R. Goodson of the College of Education, Ohio State University, has described the implications of social engineering in public school administration.¹

Dr. Edward C. Tolman of the University of California, recipient of the 1949 Kurt Lewin Award, in his Memorial

¹ Goodson, Max R. "Social Engineering in a School System," *Progressive Education*, XXVI (1949), 197-201.

Lecture^a added to the theoretical structure of the basic concepts of group dynamics in social learning. Dr. Tolman stressed the nature of the influence of drives, beliefs, goals, perceptual readiness and perceptual blindness. His concepts, methods, and findings will more adequately illuminate the complex nature of student mores.

Dr. Harold Fields, of the Board of Education of New York, recently reported in an unpublished manuscript the development of a group-interview technique used in the selection of teachers. This technique utilizes several of the common group-dynamics evaluation procedures such as systematic observations of candidate behavior in situational tests involving six candidates. This method demonstrates how individuals behave in situations of reality and is a more realistic evaluation than paper-and-pencil test material. A similar procedure is now being considered for the selection of candidates for admission to one of New York's medical colleges. Numerous professional and lay organizations are also utilizing group processes in their administrative procedures and in program development.

It is difficult to discuss research and evaluation in group dynamics when the field, as such, is not yet adequately defined, or acceptable to many in the professions related to social science. The basic philosophy, concepts, values, and skills which constitute the common core of group dynamics theory and practice will make a fundamental contribution to the improvement of human relations and the advancement of social science because this core is rooted in established and fundamental concepts of the basic social sciences.

The controversy over the nature and importance of group dynamics has been widespread. In fact, the cry of the antagonists is raised so loudly that one is tempted to echo Shakespeare, "The lady doth protest too much, methinks." The feelings of Dean Browne are shared by many critics who are concerned with: "the cult" of the proponents; "the uncritical acceptance" of group-dynamics discoveries; "the verbiage" in which they (the proponents of group dynamics) are imbedded; "the pseudo-learned special dialect"; "the wild

^aTolman, Edward C. "The Psychology of Social Learning." *Journal of Social Issues*, Supplement No. 3, Dec. 1949.

oscillation and breakdown which results from the feed-back correction and over-correction"; and "the diffuse vagueness of the literature on group dynamics." However, as Dean Browne urges, "It would be part of wisdom first to try to understand it," and since this social innovation is established upon the theory and practice of a multi-disciplinary field it behooves the critic to be reasonably well acquainted with the theory and practice of these several disciplines before he criticizes the emerging ideas, concepts, practice and research of group dynamics.

What of the nature and trends in group dynamics research and evaluation?

The importance of research in group dynamics—interpersonal relations, if you will, is no longer a rhetorical question. Human relations today produce problems of major significance in politics, industry, medicine, and community living. Education has its own personnel problems.

It is necessary, however, to recognize certain technical limitations. Let us consider, for a moment, the requirements of research which the workers in group dynamics studies have found vexing. Dr. Richard Crutchfield, writing in the *American Psychologist*¹ for the joint Committee on Public Service Standards in Social Psychological Research, reported:

In its phenomenal growth during the past fifteen years social psychology has exhibited certain faults common to any rapidly growing field of science. There has been an unevenness in the quality of the research carried on and an unevenness in the training and competence of research workers. Moreover, because its problems have an immediate bearing on practical problems of everyday life, the applications of social psychology have tended to outstrip basic research. Practical pressures will continue to favor the applied phases at the expense of basic theoretical research and methodological development upon which sound application must be founded.

Dr. Crutchfield here describes one of the reasons why there is so much confusion about group dynamics in the minds of practitioners of personnel work and of the social science world in general.

¹ In the *The American Psychologist*, IV (1949), p. 112.

Dr. Donald G. Marquis,¹⁰ in his Presidential Address to the American Psychological Association, in 1948, also stressed certain difficulties in achieving the objectives of social science in the study of group dynamics. Dr. Marquis indicated that early research in psychological frontiers suffers for a lack of theoretical structure to guide the inquiry, of an accepted and adequate terminology, of "standard measurement techniques for the relevant variables," and that it suffers because, as is true of all workers in new fields, those engaged in group-dynamics research are often "dismayed at the absence of the simplest kinds of taxonomic data on the materials of their study." Consequently, early research reports often appear to be inferior and quite unrelated.

These difficulties, described by Crutchfield and Marquis, prompt the kinds of criticisms cited by those who challenge the group-dynamics approach to an understanding of social phenomena, described earlier in the words of Dean Browne.

The *verbiage* of every professional in-group tends to exclude others temporarily. The terminology of psychoanalysis and of atomic physics were disturbing a short time ago, yet today they contribute to the language of every family. It would seem necessary, however, for us, who are interested in the objectives of group-dynamics, to guard against any exclusion by association or communication if the positive and constructive contributions this vigorous field of endeavor has to make to knowledge and skills in interpersonal relations is to be realized.

Those who are disturbed by the development of interest in group dynamics must likewise examine carefully these developmental manifestations, so that they may gain proper perspective in the examination and use of the knowledge that is constantly being developed. The immediate objectives of research in group dynamics are to develop: (a) a respectable theoretical structure to guide their inquiries; (b) an acceptable and adequate terminology; (c) standard measurement techniques for the relevant variables; and (d) the collection of adequate taxonomic data. These are difficult tasks which require time and the integration of several heretofore apparently unrelated disciplines and frames of reference. Those

¹⁰ In the *The American Psychologist*, III (1948), p. 431.

involved in the development and application of applied social science research findings concerning group dynamics either as critics or workers will have to proceed with vitality tempered with common sense and self-discipline. Recently a Dean of Students remarked:

We are establishing a faculty-student advisory program. We do not know how to group the advisees. We do not know how to introduce the available faculty advisors into the groups so that the group will achieve maximum productivity with respect to their needs. How can one form groups and include a faculty advisor when there is such a wide variance in individual capacity, experience, expectation, and in basic personality structure?

Of course, we can use the traditional unitary systems of grouping, such as intelligence or age or problem, but this is neither effective nor realistic. Heterogeneity is a conspicuous and common characteristic of group relations. It is always necessary to learn how to work with people who are different. It is naive to educate and to speculate or to rely upon the possibility of always being able to work with those who are of the same color, the same religion, the same values, and the same basic capacities. The problem of this particular personnel administrator is to help the faculty advisers and the students to learn how to work together with their differences. The problem of grouping is made difficult by a number of component problems. Dr. Morton Deutsch of the Research Center for Human Relations of the New School for Social Research, has described in detail the influence of competition and co-operation on group process and development.¹¹ Another often neglected aspect of group experiences is the psychology of learning. In this group guidance situation, learning is an important factor. Dr. Herbert Thelen, Associate Professor of Educational Psychology, University of Chicago, has made a worthy contribution to understanding this aspect of group process in a recent review of "Group Dynamics in Instruction: Principle of Least Group Size."¹² As the title indicates, this

¹¹ Deutsch, Morton, "An Experimental Study of the Effects of Cooperation and Competition Upon Group Process." *Human Relations*, II (1949), 199-231.

¹² Thelen, Herbert A., "Group Dynamics in Instruction: Principle of Least Group Size." *The School Review*, XVII (1949), 139-148.

study also treats with the importance of a desirable *number* of participants in a working group. Learning to work together is a prerequisite of group-problem solving which no idyllic platitude of brotherly love will satisfy. The hazy generalizations with which we often diagnose and prescribe for many student personnel problems are rarely specific enough to resolve the knotty problems of interpersonal relations.

A director of student activities who has responsibility for student housing states,

We have practically no social communication between the dorm students and the fraternity groups. Our campus has achieved no group standards which are commonly accepted, and one result of this lack of communication and lack of standards is a lack of morale and cohesion in the larger and common educational and developmental program. Our rules and regulations won't work.

In fact, without careful and respectable research there is no easy or fruitful answer to this problem. Personnel administrators will do well to consider seriously the types of research studies now being carried out in the area of group dynamics and interpersonal relations.

These situational problems indicate the need to understand the basic dynamics of the interpersonal relations but they also emphasize the need for a specific type of research and applied skill training in the professional training of student personnel administrators. Education and training in applied social psychology, group dynamics, and action-research skills, necessary for effective change, are too often luxury items or afterthoughts in the formulation of a program of professional education. The importance of research skills has often been minimized in the graduate preparation of our personnel administrators in higher education. If college personnel administration is to continue to be a respectable and sturdy profession more attention must be given to the development of student personnel policies and procedures substantiated by appropriate and reliable research.

"After a year and a half of study a committee of five members of Princeton University faculty released today a 7,000 word report on the state of undergraduate faculty relations at

Princeton." Reported in the Sunday *New York Times* on March 19, 1950, this article described "tensions on the campus," the need for "students and faculty to know one another better," the need for "undergraduates to forget 'the fear of apple-polishing' and to take the initiative with faculty members to recognize the obligations of a kind of campus citizenship not unlike their civic obligations in the community." It is necessary for the college personnel administrator to have at his command the methods and the skills of research which will enable him to isolate the basic problems, review the knowledge of such problems in other spheres of influence and interpersonal relations, initiate sound and revealing procedures for preliminary observations, construct a reasonable theory of causation, and verify such assumptions through the application of procedures which promise some reasonable assurance of reducing the tension.

The Phelps-Stokes Fund recently released a report¹³ of the needs of some 400 foreign students from Africa. These students came here inspired with a desire "to aid their homelands toward independent status or simply to better the lot of their fellow-countrymen." Patterns of segregation and discrimination in American colleges and universities embittered these hard-working, self-sacrificing students who came to American colleges because they offered a wider range of courses and experiences. This is a personnel problem of our native born American Negroes, Jews, Catholics and other cultural groups. To ignore or give lip service to theoretical democracy without doing something practical about the problem will contribute further to our national and international discord. This is a great opportunity, and much is being done to analyze the problem and to meet it in a concrete manner, but it is with just such problems that those involved in group dynamics studies are concerned.

Is it not an appropriate moment for college personnel administrators to seek financial support for a series of research studies of the most pressing college personnel problems which

¹³ This report was made public by the offices of the Phelps-Stokes Fund, 101 Park Avenue, and was prepared under the leadership of Dr. Ruth C. Sloan of the State Department and Ivor G. Cummings of the Colonial Office. Dr. Channing Tobias reported that the project and report were strictly private and not official in character.

are essentially problems of interpersonal relations and group behavior? Requirements of such research work demand a team of specialists in group relations, social psychology, personnel administration, mental hygiene consultants and others oriented to research procedures and also to practical problems in student personnel administration. It will be necessary to delimit many of the problem areas and to recognize that many of these pressing problems are essentially those which have to do with group relations or with the potentialities of the group as an appropriate medium for the satisfaction of certain student personnel needs.

Psychological research within our time has accomplished respectable achievements in comparative psychology, physiological psychology, in the psychology of learning, of mental abilities, and in social psychology involving political science, sociology, anthropology, and economics. The frontiers of research in human relations and of group dynamics are beyond the daily practice of most of us.

Kurt Lewin, a well-spring at the frontier of research in interpersonal relations, described in his last writing the cutting edges of research in group dynamics. Those who have carried on this work have been striving to reduce the unknown. Lewin was certain that "the scientific aspects of this development (e.g., group dynamics research) center around three objectives: (1) integrating social sciences; (2) moving from the description of social bodies to dynamic problems of changing group life; and (3) developing new instruments and techniques of social research."¹⁴

The student-personnel administrator can recognize that the student and faculty society with which he works provides a challenge for such study. Certainly the dean, adviser to students or director of student activities, as well as the psychological counselor, need at all times to remember the functional importance of: (1) the social environment and the sociology of the community in which the individual lives and works; (2) the individual differences in behavior and especially in responding to the environment; and (3) the cultural mores and standards which influence the needs of the individual and

¹⁴Lewin, Kurt. *Human Relations*, I (1947), 5.

the dynamics of the social units of the collegiate society of which the face-to-face group is a basic structure. We can utilize the basic knowledge of sociology, individual clinical psychology and cultural anthropology, but we must learn to apply this knowledge in situations of reality of which we and our students are a part. The nature of individual behavior and of the behavior of campus groups as groups is constantly changing. No complacent, static concept of the nature and function of these social units of the college will suffice. The personnel administrator must develop more critical procedures for the examination of the social phenomena of student and faculty life. The approach to such understanding utilized in the phenomenological approach to social and group experiences is indeed promising and challenging. Dr. Robert B. MacLeod,¹ Head of the Department of Psychology, Cornell University, has described certain professional needs which many personnel administrators consider of grave importance to the development of our professional skills. The first is the need for a systematic procedure of observing and describing the characteristics of experiences of people in groups. The second is the need to suspend many, if not all, of our naive assumptions as to the underlying mechanisms which prompt the behavior of people in groups. Third is the need to develop a set of principles by which it will be possible to determine what is happening in our college social and group life, how it occurs, when and where. Ultimately we may know why.

¹ Macleod, Robert B., "A Phenomenological Approach to Social Psychology," *Journal of Psychological Reviews*, LIV (1947), 193.

THE CREATION OF AN EFFECTIVE FACULTY ADVISER TRAINING PROGRAM THROUGH GROUP PROCEDURES

IRA J. GORDON

Associate Professor and Counselor, Counseling Bureau, Kansas State College,
Manhattan, Kansas

You may recall that last year, at the convention, Dean Maurice Woolf of Kansas State left us with the statement: "A blissful unawareness of the impossible is all you need," in order to venture forth on a faculty advising program. He further laid down for us some basic concepts underlying an approach to faculty cooperation. At that time I was a faculty member at Kansas State attending the sessions as an on-looker. This summer, after joining the Counseling Bureau, the author felt that there was a chance to put these concepts into effect to a degree beyond which they had been tried. So to speak, we decided to demonstrate the efficacy of the concepts, and our beliefs in the value of faculty participation in advising. Thanks to Dean Woolf's groundwork, we had a faculty advisory program, and a faculty group of 250 who were involved in it. The problem that presented itself was the utilization of these advisers so that they could function effectively at their work with freshmen. The nature of the situation was such that these people, to a great degree untrained, had to be exposed to a training program of a dynamic nature over a relatively short period of time—the Fall Semester.

Faculty advisers were spread out over virtually one-third of the staff in all of the various schools. These staff members were responsible each for a small number of freshmen, usually ranging from six to ten. These faculty members were ill-trained, and many of them were new or had had no training. This difficulty was created by an old-line feeling on "the hill" that advising required no skill; that any intelligent professor can give "good advice" to students. There was also the feeling

that college students should be mature, and should not require help. Some felt that students have no problems other than vocational choice.

The Counseling Bureau did not exert administrative control over the advisers. They were under the control and pay of the academic deans and their names were furnished to the Bureau. Therefore, there was no direct line of authority between the Bureau and the advisers. The former functioned as liaison, and as the data-supplying agency. The Bureau had, before September, 1949, attempted to provide some minimum of training, mostly through from one to four short lectures covering skills in test interpretation, concept of the counselor's role, specific information, etc. (This was rather limited in scope.)

The advisers were on the job. They had been given the cumulative folders, they were seeing students, and many of them felt that they could not make use of the information the Bureau was furnishing them. There was a strong need for holding the cooperation of the faculty gained over the last few years, and a strong need to move the program forward on more solid ground. With this in mind, the author, with the support of the Dean of Students and the Counseling Bureau staff, decided to institute a volunteer training program for the faculty, using small group procedures as the method of instruction.

Our major philosophy governing the operation of these groups was that of democratic group procedures. We desired to keep the situation free and permissive so that the groups would feel free to move along the lines they wished, and at the rate they wished. We desired that participation remain on a voluntary basis, so that those who wished to join could really feel in harmony with the program. We wanted the situation to be one in which negative and hostile feelings, personal feelings, could be expressed. We intended to place responsibility for learning in these groups where we felt it belonged—on the advisers. The training program, therefore, was built around group discussion, role playing and live experiences.

We believed, and our experience has substantiated our faith, that these groups, on their own initiative, would cover the areas that they considered the most significant to them, and

that there would not be much discrepancy between what the professional counselors considered important, and what the advisers considered important. On this belief, we did not intrude our ideas on what should be covered, or attempt to indoctrinate the participants with any one counseling point of view. We felt our role to be that of supplying resources when asked, giving aid on *how* to discuss what they had decided was pertinent, and, after the collections of people had become groups, to participate as members in the full sense of the word.

This idea was presented to the faculty advisers at a meeting on September 7, 1949. The basic philosophy behind the program was included in the presentation, as well as a partial list of the possible areas to be covered. Ninety-seven advisers, including many department heads, and all save one of the assistant Deans of the various schools on the campus volunteered to participate. A breakdown of the figures reveals the following information: 65 per cent of the advisers in the School of Home Economics attended sessions, 41 per cent of the advisers in the School of Engineering, 34 per cent of those in Agriculture and 24 per cent of those in Arts and Sciences participated in the first semester.

They were divided into training groups on the basis of the amount of time they had free, the length of time they wished to participate, the hours of the week available, and the departments in which they taught. An attempt was made to make the membership of each group a heterogeneous one, so that there could be a free exchange of ideas and information among the people with varied backgrounds and training.

Three groups, consisting of a total of thirty members, met for one session a week. Three, consisting of thirty-five members, met for one session every other week. The author was the resource person for these six groups. Two groups, totaling eighteen members, met once a week for five meetings before the advising period, and one meeting during the period. They worked with Professor Paul Torrence, Bureau Director. These people, because of time pressure, decided to meet for this short length of time. Two groups, of thirteen members, met once a month with Miss Dorothy Mitchell of the Bureau. They held a few extra meetings.

As in all situations, there were several limiting factors. It was not possible to arrange for any financial or released-time incentives for participants in the training program. Indeed, all faculty advising is "extra" work without financial compensation. The utilization of time proved to be a difficulty. Meeting times had to be arranged to suit most participants, and some who wished to join were unable to do so.

Since the College serves the entire State of Kansas, many had to miss meetings because of extension or other obligations. Faculty members are often used by state and local agencies in consulting, judging, and other roles. Many attempted to attend other group meetings to make up, but there was some loss of continuity and group unity because of this.

There were some powerful positive factors in operation that more than counterbalanced the above limitations. The faculty advisers felt that such a program was needed, many felt that they were inadequate. There was a feeling, more covert than overt, that such a program could contribute to personal growth as a teacher and as an individual.

The students, through their planning conferences, had made recommendations that faculty counseling was essential, and that advisers should be trained. The deans of the respective schools were interested in the creation of the program. All concerned displayed a strong spirit of cooperation.

One difficulty that presented itself after the program had started, and one that we had anticipated, was the normal one that arises when any group of people, competent in their own fields, are called upon to undertake new learnings and to use new procedures quite removed from their own. For example, many of the advisers have come from the physical-science and technological areas where they have long been trained in individual research, and where they have conducted classes on a lecture as well as laboratory basis. Group thinking and group processes were an essentially trying experience for many of them at the beginning. The author feels, however, that by the use of process observers, and by the resource person from the Bureau expressing these "trying" and negative feelings when they arose, that this difficulty was mostly overcome.

What did the groups discuss and do? The following is a list

of topics, discussed by at least two-third of the groups, and selected out of the protocols:

1. Test Interpretation
 - (a) Meanings of tests, test scores
 - (b) How to apply the information, interview techniques
2. Philosophy of Education
 - (a) Who should go to college?
 - (b) Responsibility of College toward student
 - (c) General education
 - (d) Curriculum construction
 - (e) Entrance requirements
3. The Problem of the Marginal Student—low aptitude, low ability, high level of aspiration
4. The Role of the Faculty Adviser
 - (a) Responsibility to institution, to student, to self
 - (b) Relationships with students
5. The Problem of increasing student contacts
 - (a) Use of social gatherings, called by adviser at home
 - (b) Use of upper-class students as group leaders of Freshmen groups (Home Economics School)
 - (c) Other schemes
 - (d) Where does responsibility for initiating contact lie?
6. Teaching Methods
 - (a) How do you teach students to accept responsibility, think critically?
 - (b) How do you create student interest?
 - (c) Grading and testing
 - (d) Group Procedures
 - (e) Student rating of the faculty
7. The Dynamics of (Student) Behavior
 - (a) Discussion of specific cases
 - (b) Role-playing-dramatized interviews

This represents only those areas covered by most of the groups. There were some groups which covered other topics, including the mental hygiene needs of instructors. On the whole, an analysis of the protocols and process charts shows a great deal of involvement in the program, many new ideas advanced, and much interchange of information among the members from the different schools.

No program would be complete without attempts to evaluate. This process is still going on. Our first thoughts on evaluation included a pre-test and post-test battery, to measure several aspects of the program. The author created, and administered to the groups, a pre-test inventory consisting of three parts.

There was no opposition on the part of the faculty after the purpose of the battery was explained, and adequate safeguards taken to insure comparative anonymity. The first part was designed to measure the individual faculty member's concept of what his role is in counseling, and his attitudes toward students. This part was a sentence-completion exercise of twenty-five items. We are still evaluating the returns on this, and the entire Bureau staff is rating the answers in an attempt to cut down on the limitations of such a projective device.

The second part was an information exercise, and the third, a miniature case study. The questions raised in the latter were revised from Strang's list in *Counseling Technics in College and Secondary Schools*, and were designed to be useful in training as well as evaluation. This was included on the theory that a utilization of knowledge in an organized, integrated fashion is necessary for effective counseling of the type done by advisers.

Only the *Sentence Completion Test* was re-administered at the end of the semester. It was felt that the other measures were too static and not valid in this type of program. The third part was used for discussion and as role-playing material in some of the groups.

In addition to this test procedure, reports of the content and process of the meetings have been kept, with a member of the group acting as process observer, using mimeographed material as a guide, and keeping a participation chart, while the resource person acted as recorder. Readings of these protocols show movement and positive changes and will be used to show growth in concept and understanding. Some recordings of role playing and discussion were made, and these will be used, too.

At the end of the semester, the author created an evaluation questionnaire that was sent to all the participants. The analysis of returns is still in process, but the evidence tends to show that:

1. We have a firm base on which to build additional training programs at Kansas State and other comparable institutions.
2. The program has had repercussion in the classroom teaching of the participants.
 - (a) Group dynamics procedures have been adapted for classroom use and experimentation in classes such

as senior mechanical engineering laboratory, freshman classes in personal health, classes in journalism, education, foods and nutrition, industrial management, and others.

- (b) A concern for the behavior of students has been met by the use of other teaching procedures.
3. Relationships with the Counseling Bureau, and use of its facilities by faculty have increased.
4. Advisers feel more adequate in their handling of test data, and have made use of the learning in interviews with students.
5. The advisers feel that they now recognize that more responsibility must rest with the student, both in counseling and in class work.

The evaluation by the faculty also shows that they gained much from the heterogeneous make-up of the groups, from the method, and from the total approach. Not all was sweetness and light, however. Of course, some faculty members, because of their own personality, or because of their long years of training, felt that such group procedures did not meet their needs. Some felt they had come for the facts, and that they did not get them presented: one, two, three. One wrote on his evaluation sheet: "I went into the program in order to have some expert counselors give me some information on counseling. . . . I was *not* interested in serving as a guinea pig for an experiment in group psychology. If you ever decide to give the advisers some pointers on counseling they can use, I shall be glad to participate." This faculty person attended only one session and withdrew. He represents an extreme minority.

Although the evaluation process is incomplete at this time, the Bureau members feel that the program has been a success. One further indication we have is that five groups are going strong in this second semester. We had decided to terminate the program before Christmas, but none of the groups decided to do so. These present groups are all meeting once a week, because we found that to be the best arrangement in our situation. We found that the groups which met each week far outdistanced the others in terms of group unity, content covered and all-round participation and satisfaction.

We in the Bureau know that we have learned much from the advisers, and have gathered many excellent suggestions

from them. We know that our knowledge of group procedures has grown greatly from the experience. We have learned much about our role in such groups, and about the faculty expectations of such a role. We are now using that learning in the spring groups. Perhaps it would be more exact to call this a "cooperative learning program" rather than a faculty training program.

We feel that the use of the knowledge of small group dynamics in creating and operating a large-scale training program for advisers is practical and successful, and that it can be applied effectively in other institutions. We believe such a program rests upon the extension of the application of personnel techniques by the counselor to the faculty. If the counselor respects his faculty colleagues, works with them in a democratic fashion, and attempts to meet their needs, he can secure faculty cooperation and participation in advising and training.

A GENETIC STUDY OF SOCIALITY PATTERNS OF COLLEGE WOMEN

DAVID S. BRODY
Montana State University

Introduction

THE present research represents an exploratory study of some of the underlying factors determining sociality patterns of 140 freshman college women living together in a dormitory residence at Montana State University during the academic year 1948-49. Sociometric data employed at the residence halls in reassigning roommates after a period of six months were utilized as a criterion of sociality. Each girl was asked to list the names of all girls in the dormitory she would like to have as roommates as well as the names of girls she did not desire as roommates. Since the girls knew that the data would be actually used for room assignments, maximum cooperation was secured. During the first week of the Spring Quarter, after the girls had moved to their new rooms, they were asked to rate the other girls in the dormitory on three traits: leadership, social qualities, and work habits. In addition, each girl filled out an inventory indicating the extent of participation in various home duties and in individual and group activities prior to her entry in college. She also filled out a Questionnaire pertaining to the parents' attitudes and their supervision of activities prior to her entry into college. Data on the *Minnesota Multiphasic Personality Inventory*, which was administered at the beginning of the academic year, were also utilized in the study.

Additional data on participation in individual and group activities in college and a measure of student satisfaction with college life were secured toward the end of Spring Quarter. However, these data have not as yet been analyzed.

Results on Item Analyses

The initial step in the analysis of data consisted of tabulating the number of times each girl was accepted as a roommate. The number of acceptances for each girl ranged from 1 to 27 with a mean of 7.6 choices.

On the basis of this tabulation, two groups of 30 girls, each representing approximately the lowest and highest 20 per cent in the distribution, were isolated.

For purposes of discussion these groups will be referred to as the "Unaccepted" and "Accepted" groups. Each girl in the unaccepted group received four or less choices as a roommate and each girl in the accepted group received ten or more choices.

Employing these two groups, an analysis was made of each of the items on the Inventory and Questionnaire. It was found that a number of items did differentiate between the two groups at the 5 per cent level of probability or better. Altogether, a total of 241 items were employed in this exploratory study and approximately 20 per cent were found to be significant.

Items included in the Inventory indicating the extent of participation in various activities prior to entry in college were classified under three headings:

- Part I. Participation in individual and informal group activities.
- Part II. Participation in home duties.
- Part III. Participation in formally organized group activities.

Each item in Part I and Part II was checked by the student in terms of frequency. (For purposes of item analysis, three categories were employed—namely, *none or little*, *some*, and *much or very much*.) In Part III pertaining to formally organized group activities, four categories were employed:

- (a) no participation
- (b) member in name only
- (c) participating member
- (d) officer or committee chairman

Analyses were made separately for each category within an item.

In Part I, the following seven items showed significantly greater participation on the part of the accepted group:

- (a) attending movies
- (b) swimming
- (c) going out on dates
- (d) touchball
- (e) hiking
- (f) social dancing
- (g) visiting friends

Significantly greater participation on the part of the unaccepted group was shown by the following two items:

- (a) playing checkers or chess
- (b) reading

In Part II, all of the significant items showed greater participation on the part of the accepted group. These items are:

- (a) selected new clothes for myself
- (b) laundered
- (c) made my own bed and straightened out my room
- (d) painted (furniture, walls, etc.)
- (e) canned fruits and vegetables
- (f) cleaned house
- (g) washed and wiped dishes
- (h) chores around barns
- (i) worked in fields (ploughing, sowing and harvesting)

Similarly in Part III, the items yielding significant differentiation showed more participation for the accepted group. These items are:

- (a) student government
- (b) high school fraternity or sorority
- (c) school athletic team

It should be emphasized that a number of other items showed consistent differentiation between the unaccepted and accepted groups for each of the categories, but fell somewhat short of meeting the 5 per cent level of significance. (Data for these items will be presented in a subsequent paper.) There would appear to be an important difference in the extent of home responsibilities between the two groups of girls. In general, girls in the accepted group reported that they fulfilled home

responsibilities to a much greater extent than girls in the inaccepted group.

The Questionnaire on Family Background included items designed to indicate parents' attitudes and their supervision of activities prior to entry in college. The first section of this Questionnaire consisted of 42 items, twenty-one of which pertained to the father's attitudes and the remaining twenty-one to the comparable attitudes of the mother.

Of this group of items, a significantly greater proportion of the accepted group indicated that:

- (a) My father provided me with a regular allowance
- (b) While attending high school, my father expected me to participate in social activities

Whereas, more of the unaccepted group indicated that:

- (a) My mother expected me to work for pay outside the home
- (b) My mother tried to push me ahead and to make me excel
- (c) My mother emphasized the importance of good manners
- (d) My father selected clothes and other personal articles for me so I wouldn't make mistakes

Included in this section were 13 additional items measuring parent-child and sibling rapport. These items were adapted in part from Terman's study¹ on the prediction of marital happiness. Girls in the unaccepted group indicated a significantly greater degree of conflict both with their fathers and with their mothers than did girls in the accepted group. They likewise showed a greater amount of conflict with their brothers and sisters.

Another series of items on family background was designed to indicate the type of control exercised by the parents relative to 21 different areas of activities. The students were asked to check the type of control exercised by the mother and father separately.

The items appear to indicate less stringent control for girls in the accepted group. However, before generalizations can be drawn from the data, analyses in terms of weighted scores

¹Terman, Lewis M. and Others. *Psychological Factors in Marital Happiness*, New York, McGraw-Hill Book Company, 1938.

are indicated. This step has not yet been taken. Preliminary analyses indicate that the significant areas are:

- (a) choice of friends of the opposite sex
- (b) going out on dates
- (c) time of coming home from dates or parties
- (d) studying school lessons
- (e) cleaning my room and taking care of personal possessions

The last section of items pertaining to family background dealt with the extent of agreement between the student and her father, and between the student and her mother, on the type of supervision of the same 21 activities listed under parental control. Girls in the accepted group showed generally greater agreement with their parents than did those in the unaccepted group.

In summarizing the data on item analysis, it is significant to note that certain items appear to yield consistent differentiation between the unaccepted and accepted groups regardless of the context in which they are found. For example, items concerning home duties, especially those involving personal responsibilities, differentiated between the accepted and unaccepted groups in the activity inventories, in the Questionnaires on the parents attitudes and their supervision of activities, and in the Questionnaire designed to measure extent of agreement between child and parent. Likewise, items concerning association with the opposite sex differentiated between the two groups of girls on the activity inventories and on the Questionnaires.

Results on Ratings of Leadership, Work Habits, and Social Qualities

As was indicated earlier in the paper, each girl was asked to rate all the other girls in the dormitory on leadership, work habits, and social qualities. The students were instructed to place a check mark in front of a girl's name and go on to the next name if they had had no opportunity to observe that girl or did not know her well enough to rate her.

Each rating was weighted from 1 to 5, 1 representing the lowest rating, and 5 the highest. Ratings for each girl were

tabulated and the mean of all the ratings she received was computed. The mean rating represented her score on the particular trait. For each of the three traits the distributions of mean ratings were symmetrically distributed and approximated normality. The girls were rated most frequently on social qualities and least frequently on work habits. The number of girls who were not sufficiently well known to be rated on social qualities was 24 or 17 per cent, on leadership 31 or 22 per cent, and on work habits 59 or 42 per cent. Thus, the girls as a group felt that they were least able to evaluate others on the basis of work habits and most able to evaluate others on the basis of social qualities.

The number of acceptances each girl received was correlated with the mean rating on each of the three traits. The highest correlation with acceptance scores was obtained for social qualities. This correlation was .59. The correlation with leadership was .52 and with work habits, .20. Ratings for social qualities and leadership are certainly significantly related to acceptance scores, but it is obvious that social qualities and traits of leadership are by no means the only factors determining acceptability. The possession of good work habits is apparently of minor importance in the selection of roommates.

When the correlations were computed between work habits and ratings on leadership and on social qualities, they were found to be .49 and .32 respectively. Thus, we find that leadership correlates almost as highly with work habits as with acceptability.

Although there is a significant relationship between work habits and social qualities, it is considerably lower than between work habits and leadership. However, leadership and social qualities are highly related to each other as evidenced by a correlation of .85. It can be hypothesized that social qualities are an important determinant in the selection of leaders among this population, but that work habits constitute another variable which is significant.

Results on Minnesota Multiphasic Personality Inventory

Since all members of the freshman class were given the MMPI at the beginning of the Fall Quarter, 1948, as part of the

Orientation Week Testing Program, it was decided to utilize these records to determine whether the accepted and unaccepted groups could be differentiated in terms of this personality inventory.

In the present study, 28 MMPI records were available on the unaccepted group and 30 on the accepted group. On the keys pertaining to the specific personality variables, it was found that the mean scores of the unaccepted group were consistently higher than those for the accepted group, with the exception of the mean score on the hypomania scale. When the 't' values were calculated, only one was found to be significant at the 5 per cent level of probability or better; this was the scale pertaining to schizophrenia. Since the schizophrenia scale is theoretically measuring withdrawal tendencies, it is the one on which we might have expected to attain the maximum differentiation.

When we compare the standard deviations, we find that the unaccepted group is consistently more variable than the accepted group on all the scales except depression. When these differences in variability were tested for significance by the calculation of the F ratio, the psychopathic deviate, the schizophrenia and hypomania scales showed significant differences in variability at the 5 per cent level or better.

The consistency of these differences in means and standard deviations suggests that the members of the unaccepted group are more likely to have personality disturbances than the members of the accepted group. The greater variability of scores for the unaccepted group is a reflection of the larger number of deviant scores in the direction of abnormality.

Differences in variability on the L, F, and K scales were all significant at the 1 per cent level or better; the unaccepted group being more variable on the F & K scales and the accepted group more variable on the L scale. Differences in mean scores on these three scales were not significant.

Summary and Conclusions

The data which have been presented in this paper suggest that the student's experiences within her family group and the pattern of her activities prior to entry in college are important determinants of her social acceptability. For example, girls in

the accepted group indicated significantly greater participation in home duties, especially those involving personal responsibilities. In a comparison of activity participation outside the home, girls in the accepted group showed more participation in social activities whereas girls in the unaccepted group showed more frequent participation in relatively solitary activities. The evidence also points to the fact that the parents of the unaccepted group tended to be overprotective and to discourage the development of independence. The girls in the accepted group felt that their parents encouraged social development to a greater degree than was true of the girls in the unaccepted group.

Apparently girls in the accepted group came from homes in which there was less conflict and greater harmony than was true of the homes of girls in the unaccepted group. The results on the MMPI suggest that girls in the unaccepted group evidence a greater tendency toward abnormality.

Reilly and Robinson² in their study of popularity among college women point out the importance to counselors of obtaining some index of the probable social acceptance of an entering freshman. Their report shows that the usual college entrance data are relatively ineffective for predicting social acceptability. Of interest is their recommendation that academic census data need to be supplemented by more vital statistics from the adolescent world. Certainly, the present study points to the value of this approach. For the personnel worker it means that if he is to understand the dynamic factors underlying social behavior at the college level, he must orient his thinking in terms of the developmental history of the individual.

²Reilly, J. W. and Robinson, F. P. "Studies of Popularity in College: I—Can Popularity of Freshmen be predicted?" *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, VII (1947), 67-72.

HOW TO GO ABOUT THE PROCESS OF EVALUATING STUDENT PERSONNEL WORK

WILLIAM M. GILBERT

Director, Student Counseling Bureau, University of Illinois

THE title of this paper is somewhat misleading and needs to be clarified. The title implies that there is a specific process, that student personnel services can be neatly defined, that there are perfectly valid criteria for determining the effectiveness and efficiency of these services and, finally, that the necessary formula for going about the process of evaluation can and will be supplied in cook-book fashion. Unfortunately, not one of these implications is justified. There is no one most-desirable way of going about the evaluation process. Student personnel services cannot be defined at all neatly; there are no criteria known to be perfectly valid and I have no secret formulas.

With these few positive statements I should possibly end this paper dramatically and sit down. However, student personnel services will not continue to be accepted on faith indefinitely. Eventually some discerning college or university administrator will rightly ask: Just what are the purposes of student personnel service and what is the evidence that these goals are being attained, or how can we get this evidence? We could not avoid the issue even if we wanted to.

Mr. Blaesser, in his statesmanlike and visionary address of last year, and after reviewing the various attempts at over-all evaluation of student personnel programs, faced the issue squarely. He emphasized: "This means a total institutional study of the needs of the students coming to the institution, and the evaluation of the outcomes of the total educational experiences at the institution."

It was rightly explained by Mr. Blaesser that it would be a "Shangri-la" university where such far-reaching, highly cooperative and expensive over-all and long-range evaluation of higher education could be carried on.

In the meantime, before such a university evolves, the college or university administrator is still going to have to allot funds to the various student personnel agencies and he is still going to want to know what evidence there is that the different objectives are being reached. He will probably not insist on perfectly valid evidence because he will be one of the first to recognize that perfection can be aimed at but cannot be expected in broad educational endeavors. And we, as personnel workers who are sincerely interested in the welfare of students, will want to know how effectively and how efficiently we are serving their welfare.

It is not possible or polite for me to make judgmental statements about the different colleges and universities you represent. However, I am sure it will not be held against me by President Stoddard and Provost Griffith if I make the simple observation that while the University of Illinois is one of the great universities, it is certainly not a "Shangri-la" university. We have our problems too, as most of the rest of you do. There appear to be good spots and not-so-good spots in our over-all student personnel program. Most of you probably have what appear to be good spots and not-so-good spots in your programs too. The desirability of some type of general evaluation is probably quite clear. One of the problems is how one should go about this process. Perhaps, by discussing some of the procedures which have been used at Illinois and some of the plans and hopes we have, ideas for developing evaluation procedures which will fit your own local situation may occur to you. Conversely, any suggestions you have for us will be deeply welcomed.

One of the first problems to be faced both chronologically and in terms of importance is that of securing general, grass-roots acceptance of any type of evaluative procedure.

In most colleges and universities, evaluation immediately poses a number of serious problems which must be faced. When we evaluate counseling services, when we evaluate registration procedures, when we evaluate health services, and when we evaluate instructional services, we are evaluating not simply services, but, perhaps even more importantly, we are evaluating the persons who are responsible for such services and the

persons who perform the services. As personnel workers we should probably be the first to recognize that many of our student personnel are not as good as they should be and that any evaluation of them immediately can serve as a threat to the individuals concerned.

Several years ago the Student Counseling Bureau at the University of Illinois conducted a questionnaire study of student attitudes regarding the effectiveness of the counseling services they received. This Questionnaire, which went out to some 3000 students, was devised by members of the full-time psychological staff with full consideration given to suggestions made by the trained part-time faculty counselors who are a part of the Bureau staff. Nevertheless, faint rumblings of concern came to my ears, and since the purpose of the investigation was not that of evaluating individual counselors, but the program as a whole, the counselors were reassured that there would be no individual breakdown of the findings. Nor was there.

Just a year ago, in response to the recommendations of a committee appointed to study the problem of the recognition of faculty counseling, the Bureau was given the responsibility and privilege of making formal recommendations, as to increases in regular academic salary and rank for Faculty Counselors. The exact statements in Provost Griffith's letter of March 23, 1949, are sufficiently noteworthy to deserve quotation:

This whole problem has been studied recently by a special committee appointed for the purpose. I am approving the recommendations of this committee as follows:

1. *Policy.* A positive program of counseling services to students based on the best clinical and guidance practices has become and should remain an integral part of the educational experiences we offer to students. The persons who do this type of work well should be rewarded for it and advanced in rank and salary in proportion to their excellence.
4. *Rank and Salary.* Recommendations for changes in rank and salary of personnel listed in the budget of the student Counseling Bureau, insofar as counseling services are concerned, will originate with the Director of the Bureau and college offices to the general administration.

These significant forward steps in student personnel practice seemed to deserve a very careful consideration of any recommendation for increases in rank and salary that would be made. Consequently, within the past several months the problem of improving the Director's evaluation of the counseling effectiveness of individual counselors was presented to the group. It was decided that an Evaluation Committee should be elected consisting of two of the faculty counselors and two of the central staff members. The general theoretical and practical problems of evaluating the effectiveness of counseling were considered democratically but briefly at a general staff meeting. The Evaluation Committee then went to work and presented a series of recommended evaluation procedures. These were then discussed at another general staff meeting.

The next step consisted in having the counselors check the evaluation procedures which had been recommended by their Committee. Their checked lists were sent in without signature. I should like to report some of the conclusions and recommendations of the evaluation committee:

- I. The committee members agreed on the following statements as starting points affecting all recommendations on specific methods:
 1. Though various attempts to evaluate counseling have been reported in the literature, the validity of no method has been established.
 2. No single method should be used as the sole basis of evaluating counseling.
 3. Every method used should be on a trial basis.
 4. Training, supervision, and evaluation are inseparable.
 5. Outcomes of whatever methods are used to evaluate counseling may serve as guides to further training of counselors.
 6. We do not feel it necessary to recommend specifically such obvious, continuous, and informal procedures as evaluating counselors for regularity and dependability in attending to duties, cooperation in the work of the Bureau, private consultation with the Director, participation in staff discussion, research, and performance of special duties such as taking part on the staff programs, work on committees, and the like. We feel that our assignment is to suggest more formal, specific, objective, special-occasion procedures to supplement these informal ones.
 7. Morale of the staff and, therefore, of each Counselor

is a prime consideration in the selection and application of procedures.

8. Each staff member should feel free to submit additional evidence (such as recordings, additional participation in fake interviews, etc.) in his own behalf and beyond whatever evidence would otherwise be used in evaluating his work.
- II. We recommend that the present methods of evaluation by the Director be continued, and that the staff consider additional methods as possible supplements.
- III. We recommend to the staff for consideration:
 1. Intake conferences. Staff members would meet in small groups on Wednesdays when no meetings of the entire staff are scheduled. A central staff member would lead the group. Staff members would summarize their work since the previous meeting. Specific problems could be brought before the group for discussion. The Director would divide his time among the groups.
 2. A survey of clients by mail questionnaire. This should cover each counselor's entire client group for a given semester, with the client anonymous and the counselor identified on the outgoing questionnaire. We would suggest that a committee be appointed to make the Questionnaire and that the committee include those staff members who worked on the similar questionnaire used previously.

These recommendations received the unanimous approval of all Counselors who then suggested that the Questionnaire survey of several years ago be analyzed further to determine whether the type of Questionnaire used would actually discriminate between different counselors.

These few examples indicate the importance of securing the acceptance of the persons involved in any evaluation procedure and the importance of attempting to minimize any feelings of threat which evaluation would involve. It is assumed that not all threatening aspects of an evaluation procedure can be eliminated completely. If one attempts to eliminate all possibility of threat, then it is probable that a *laissez faire* policy will ensue which will result in no progress.

Even though the evaluation of individual counselors in a counseling bureau presents the issue of securing the acceptance of evaluation in its most critical form, it is still a considerable distance removed from the general goal of securing acceptance for an over-all evaluation of all student personnel agencies. At

least a tentative acceptance of the desirability of an over-all evaluative procedure by the persons and agencies who would be evaluated would be desirable even before the appointment of an evaluation committee such as that suggested in Mr. Blaesser's address of last year. It would be possible, of course, for some interested agency, such as the Counseling Bureau, which has already carried on a self-evaluative procedure, to recommend to the higher administration that such a committee be appointed. If such a recommendation were acted upon favorably, in the absence of prior consultation with the various student personnel services, it seems possible that unnecessary protests and eventual lack of real cooperation from some of the agencies would be the result.

At Illinois this second step in evaluating student personnel work, that is, the appointment of an over-all evaluation committee, will probably be approached in a somewhat different manner. In connection with the authority given the counseling bureau to make recommendations with respect to increases in rank and salary for faculty counselors there was also appointed, at the request of the Bureau, an advisory council. I quote from the Provost's letter again:

An Advisory Council to the Director of the Student Counseling Bureau is authorized, this Council to be composed of a representative of each college and school. Membership in the Council will be on a revolving basis with members appointed for three-year terms. For the first year, the one-year and the two-year and the three-year appointees shall be determined by lot. A vacancy will be filled by a staff member from the college or school which loses a representative on account of the rotating membership.

In order to set up this Advisory Council, I should appreciate having an early nomination from each dean and director.

This Council has been meeting with the Director of the Counseling Bureau each month during the present school year. One of the main problems which has been considered by the Council is the effectiveness of the various college registration advisory systems. These advisory systems do not appear to be of equal effectiveness in all colleges, a condition which has resulted in the publication from time to time of critical editorials in the school paper. As a result, the Advisory Council recom-

mended to the Director that questionnaire appraisal be made of the various advisory systems with questionnaires being sent to the students affected, the advisors, and to the academic deans. The remainder of this paper will consist of a description of plans and hopes which the Director of the Counseling Bureau now has.

It is hoped that before any evaluation of the *advisory systems* is actually put into effect it will be possible to secure approval for an *over-all* evaluation of student personnel services. Specifically, it is hoped that it may be possible to secure the adoption of both the general evaluative procedure suggested by Dr. Kamm and Dr. Wrenn, and of the one which Dr. Kamm will describe to you later today. Securing the adoption of these procedures or modifications of them will possibly not be an easy task. It is one which probably can be accomplished, however, provided the various individuals concerned have time to consider the proposals and are given the opportunity of making suggestions regarding them.

The next step would be to recommend to the higher administration that an over-all Evaluation Committee be appointed. This Evaluation Committee should probably consist of representatives of all of the various colleges and schools as well as representatives from all of the different student personnel agencies including the Dean of Student's Office, the Office of Admissions, the Health Service, the University Union which carries on a broad program of student activities, the Speech Clinic, the Housing Division, and the Placement Bureau, and possibly student representatives.

The third step in going about the process of evaluation would naturally follow from this second step. It would seem that the first task of the Evaluation Committee would be to discuss the results of the over-all, general evaluation of student personnel services and then to proceed to the problem of making a more detailed evaluative study of those services which appeared to be most in need of strengthening. The whole problem of criteria of effectiveness and efficiency of student personnel services would probably concern this Committee for some time. Since Dr. Strang will probably discuss with you the limitations of various criteria on the basis of which student personnel services

might be evaluated, I will not attempt to examine these questions with you. It seems probable that the list of criteria suggested in the revised brochure "The Student Personnel Point of View" published by Dean Williamson's Committee under the auspices of the American Council on Education, as well as other more specific criteria, such as those suggested by Dr. Aiken in his report to the Fourth Annual National Conference on Higher Education in April of last year, would be considered.

It might be reasonably expected that the Evaluation Committee, after considering the various criteria and various methods of procedure which could be used, would refer the problem to representatives of each of the different student personnel services for further consideration and for recommendations as to specific methods and procedures of carrying on an evaluation program in their own agencies.

While there are many possible objections to a Questionnaire type of appraisal of student personnel service it is one of the few practical and not prohibitively expensive means for securing some rough estimate of the apparent value of the service. There is one point in connection with Questionnaire surveys which I feel has not been adequately emphasized and that is that a student's responses to a Questionnaire will necessarily be influenced by the knowledge which the student possesses, not only of the services which are actually available but of those which theoretically could be made available. Thus, as part of a Questionnaire appraisal of any given service it would seem advisable to supply the student with a description of what services might reasonably be expected from any given type of agency.

From one point of view, at least, it may be fortunate that students are not more aware than they are of some of our more specifically stated objectives. It might be of considerable interest, for example, to submit to a representative group of students in any of our colleges and universities the eleven objectives of general education recommended in the report of the President's Commission and to have the students indicate on a simple scale the degree to which they felt their general college education already had, or seemed to be, in the process of helping them to reach these goals. The results could be startling.

After the more specific evaluation proposals of the different

student personnel services had been referred to the Evaluation Committee for approval, and after the evaluations had been carried out, the fourth step in the process of evaluation would then confront this Committee. This fourth step would consist in the Evaluation Committee's carefully examining and discussing the results of the detailed evaluation of each agency and of their arriving at a series of specific recommendations which would be automatically transmitted to the Director or person in charge of the specific student personnel agency. Such a series of recommendations should, of course, be influenced by a functional analysis of all student personnel services with the view of expanding those services which needed expanding and of contracting those which seemed to be over-expanded. This, as most of you will recognize at once, is one of the most difficult, and delicate, and perhaps even dangerous steps in the whole process of evaluation. It is my own experience that practically every Director of any student personnel service is firmly convinced that his service would improve immeasurably if it were only expanded. This suggests, of course, that the Chairman of the Evaluation Committee should be a person of the greatest possible diplomacy, wisdom, and ruggedness. In addition it would be highly desirable if the college or university Administration would be able to indicate, within some fairly definite range, at least, the total amount of funds which might reasonably be expended for all student personnel services. It seems possible that a wire recording of the proceedings of the Evaluating Committee at this point could provide some valuable research material for determining the extent to which the leaders of various personnel services were actually interested only in the welfare of students.

The next step in the total evaluation process would consist of repeated and improved evaluations of the various personnel services at regular intervals. This should prove to be a relatively easy task if the other steps in the process already mentioned have been successfully negotiated.

The final step in going about the process of evaluation might then consist of an over-all basic evaluation of the outcomes of higher education including instruction. At this point, the Evaluation Committee would probably have to be enlarged to include

other standing Committees in the university such as the Educational Policy Committees, the Admissions Committee, and others. It seems that any university which has actually reached this stage in the evaluation process should have little difficulty in securing the large funds necessary for the over-all evaluation of their educational program from any one of the national organizations which would subsidize research. In addition, that college or university should be placed at the top of some role of honor which would be devised by the American College Personnel Association.

What has been said can be summarized in a few sentences. The way to go about the process of evaluating student personnel services is to take account of what we know about people in general and to make full use of good democratic administrative procedures at every step in the process. If the process of evaluating student personnel services cannot be carried on in this fashion a very critical examination of the whole basic structure and functioning of the college or university itself needs to be accomplished first.

MAJOR LIMITATIONS IN CURRENT EVALUATION STUDIES

RUTH STRANG

Professor of Education, Teachers College, Columbia University

EVALUATION is a complicated business. It necessitates (1) clarifying goals or objectives; (2) devising methods and instruments for securing evidence that each of these specific objectives has or has not been attained; (3) gaining information about the changes that have taken place in individuals, groups, or community; and (4) passing judgment on the "goodness" of the changes. An excellent review of the literature was published in January, 1949, by Froehlich.¹

The evaluation of evaluation is still more difficult. This is because there are so many kinds of end results and processes to be evaluated—the personnel program as a whole, the adequacy of staff, the provision of certain services, the processes of counseling and of group work. Moreover, these are evaluated for different purposes and on different levels of scientific precision. For example, a teacher may use information-evaluation methods, such as obtaining from students a simple written statement regarding the effectiveness of his teaching or holding a group discussion of the methods used in the course. These suggestions for improving his teaching may be very useful in modifying instruction for the better even though they meet few of the criteria of scientific evaluation. The effective teacher continuously studies his students' progress toward the definite goals in the course.

Despite its difficulty, evaluation of personnel work is necessary if the college personnel officer is to maintain his status. Administrators, the general public and students want to see results; they demand proof of the effectiveness of counseling and group work.

¹ Clifford P. Froehlich. *Evaluating Guidance Procedures*. Washington, D. C.: Federal Security Agency, Office of Education, 1949.

With the increased interest in evaluation in every area of education, methods of evaluation of personnel work have been improved. But because of the difficulty and complexity of ascertaining changes produced by student personnel procedures, there are still major limitations in current evaluation studies—in surveys of the program as a whole, in evaluation of different services, in appraising various kinds of counseling and psychotherapy, and in the evaluation of group work procedures.

Surveys of the Personnel Program

Surveys of personnel programs tend to be either anecdotal or atomistic. The anecdotal type are valuable in giving glimpses of present practice which can be appraised theoretically. They fall short of adequate evaluation in being somewhat subjective—the investigator may select the aspects that appeal especially to him; if his mind-set is critical, he will focus on the unfavorable procedures; if his mind-set is favorable, he is likely to note the incidents that will create a good impression. Almost everyone has an unconscious bias that is difficult to recognize and control.

The detailed lists of criteria on administrative leadership, provisions and facilities for guidance, and in-service education; on the preparation and qualifications of the guidance staff, their growth in service; the specialized services available; the guidance and informational services available to students; the counseling and placement services; follow-up studies; relation of guidance to curriculum and instruction; use of community resources—this detailed analysis of the program is very useful in calling attention to the possible scope of the program and to standards in training and performance. It falls short of effective evaluation in three important respects:

1. It is too atomistic—it considers each item separately without much attention to its relative importance and relation to other items. For example, in a college in which the faculty-student load was very small, the faculty members were selected with reference to their qualifications for counseling, and the faculty adviser was the key person in the guidance program, the need for special personnel workers would be quite different from that in a college having a traditional subject-centered faculty.

2. The qualitative aspect is neglected. In two colleges, both reporting individual interviews with students, one might have interviews of a high quality, while the interviews in the other institution might be perfunctory and even detrimental. Similarly, autobiographies might be used in one college to help students to gain self-understanding, and in another college they might increase the students' insecurity and anxiety. In one college the cumulative records might be kept up to date and used much more effectively than in another institution. The check list or scale type of evaluation does not supply data on the important qualitative aspects.

3. The effect of the qualifications and services on the students is not known; in other words the crucial question of evaluation is not answered, namely, "Do the procedures we believe to be effective really make desirable changes in students, in groups, and in the community?"

In studying the personnel work in a college, little progress has been made in defining concretely the changes that should result from an effective personnel program. Last year at the annual convention, one large group pooled their opinions on this subject and listed specific changes in students' behavior and attitudes, faculty cooperation, group activities, and in the community, which they thought should be the outcome of personnel work.

Evaluation of Different Services

Educational and vocational guidance are two services that have most frequently been subjected to evaluation. Much dissatisfaction has been expressed regarding the usual criteria of success of vocational guidance—number of positions held, length of time positions were held, reasons why person left the position, reports by employer of worker's proficiency and job satisfaction of worker. Obviously, a combination of these criteria is more satisfactory than any single item. In his evaluation of the State Consultation Service at Richmond, Virginia, Froehlich³ moved toward a more adequate combination of criteria—criteria of occupational adjustment and personal adjustment, the client's attitude toward the counseling service and change

³ Clifford P. Froehlich. "Toward More Adequate Criteria of Counseling Effectiveness." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, IX (1949), 255-67.

in occupation, and his preparation for the job. Admirable as this effort is to obtain the most accurate opinions and to apply statistical methods as a test of the reliability and validity of the ratings, it has certain important limitations, clearly recognized by the investigator:

1. The agreement between the interviewer's and counselor's rating for occupational adjustment was not as high as desired.
2. Some of the questions are ones on which the client would not be expected to have much basis for judgment, such as the relative value of different counseling procedures, especially as the client's attention was not focused on the process.
3. The interviewer's basis for rating the client's adjustment was meager.
4. Much more information is needed about the individual's capacity and the environmental conditions that might make vocational and personal adjustment either easy or difficult for him, overriding, as it were, the effect of the counseling service per se.

A much more specialized aspect of evaluation of the college advisory system is to be presented at this meeting by Friedenberg. This represents an ingenious and detailed attempt to have the recipients of the service evaluate faculty advisers. From such an evaluation the faculty adviser can obtain many helpful suggestions for the improvement of his services. It clarifies the areas in which the faculty adviser can best work, and indicates the need for specialized services. The same limitation as was mentioned in the preceding study holds here, namely, the students' inadequate basis for evaluating a process in which they have had so little background of experience or study. However, the concrete cases do give the student an opportunity to focus attention objectively on the counseling process. After having obtained this information the problem of appraisal is still unsolved: Who is right—the student or the person who has studied counseling and psychotherapy?

Evaluation of Psychotherapeutic Counseling

Considerable work has been done on evaluating the non-directive interview. Much of this has been along the line of showing increased insight on the part of the client as the inter-

views continue. The assumption is that insights expressed in the interview are in themselves evidences of adjustment and will affect life adjustment. This assumption has been questioned. Consequently, evidence of adjustment in life situations over a long period of time has been considered the only valid measure of the success of the psychotherapeutic interview.

Even this criterion has its limitations insofar as environmental conditions may be so destructive as to prevent the good adjustment that might have taken place under ordinary conditions. Another limitation is the lack of evidence of the individual's initial capacity for adjustment. If the client's problem is deep seated, persistent, and pervasive, failure to show much progress could not be attributed to poor counseling techniques.

Evaluation of Group Work Procedures

As in the evaluation of interviews, too much reliance has been placed on subjective evaluation of the group work process. Some recent studies, however, have obtained reports from the participants themselves and from those who have had an opportunity to observe them some months later. For example, Lippitt³ obtained evidence of actual change in the performance of leaders who had spent two weeks in a workshop that featured group discussion, role-playing in sociodrama, and interviewing. Both outside observers and the members of the workshop reported that because of the workshop they were able to do more effective work with their community groups.

The College Evaluation Officer

A new position seems to be emerging in colleges and universities. This is the college evaluation officer, with training in measurement and evaluation. This work is closely related to, and has often grown out of, the research function of the personnel department. Such an officer was described by Findley in a meeting of the American Educational Research Association. This officer would render valuable advisory service to the faculty in defining objectives, developing instruments to measure them, assisting in the collection of data, and appraising and interpreting the information collected.

³ Ronald Lippitt. *Training in Community Relations; a Research Exploration Toward New Group Skills*. New York: Harper and Brothers, 1949.

Summary

The major limitations in evaluation studies seem to be:

1. Failure to define the outcomes of personnel work concretely as desirable measurable changes in students, faculty members, groups, and community.

2. A too narrow approach instead of a comprehensive study. All of the approaches that have been used in evaluating guidance procedures have some value. We need to know about the staff and the procedures being employed; student opinion and expert opinion as to the effectiveness of the procedures are helpful; follow-up studies supply essential information on life adjustment. The intensive study of specific techniques and the control-group and within-group experimental methods also contribute to our understanding of the effectiveness of student personnel work

3. Mass rather than individual treatment of the data collected. Instead of studying the data collected as a group, an appraisal of each student should be made individually in the light of his previous progress. This is the case-study approach to evaluation. It seems to be the only adequate way to appraise changes in students. It enables the investigator to take into account the student's capacity for adjustment to college and environmental conditions that may be reinforcing or defeating the college personnel program. A case study is made of each student; these records are studied individually and a judgment made of the student's social, emotional, physical, and intellectual development. These judgments may then be treated statistically and checked as to reliability and validity. In the case study approach to evaluation the service and the research functions of student personnel work come together; one reinforces the other.

AN INVENTORY OF STUDENT REACTION TO STUDENT PERSONNEL SERVICES

ROBERT B. KAMM

Dean of Students, Drake University

Introduction

INCREASINGLY, we are becoming aware of the need for evaluation of our student personnel programs. Now that the peak veteran enrollment has passed and we are faced with somewhat declining enrollments and the corresponding reduction in income, we need, all the more, to be able to take stock of the quality of our services.

Just a year ago, considerable time was spent at this convention in a discussion of the evaluation of student personnel services. Dean Willard W. Blaesser, then of Washington State College and now with the United States Office of Education, spoke on the subject "The College Administrator Evaluates Student Personnel Work" (1). Dr. John H. Rohrer, Professor of Psychology at the University of Oklahoma, presented a paper entitled "An Evaluation of College Personnel Work in Terms of Current Research on Interpersonal Relationships" (4).

A comprehensive review of the literature dealing with evaluation was presented by Blaesser. Reference was made to such studies as those of Hopkins (3) in 1925, Brumbaugh and Smith (2) in 1930, Williamson and Sarbin (5) in 1940, as well as others. As the title of his paper indicates, Blaesser dealt with evaluation from the point of view of the administrator.

But what about the student? Does he think what we have to offer is of value? Are our various services really functional in his college experience? Are we supplying those services which really meet his needs? How about securing "consumer reaction" to our student personnel services?

The above, and other questions, were asked last year at this convention. In fact, there was so much interest in the general subject of evaluation that the Program Committee has again

seen fit to provide a session in which the problem may be discussed.

A Student Reaction Form.—For some two or three years now, Dr. C. Gilbert Wrenn, Professor of Educational Psychology at the University of Minnesota, and I have been experimenting with a student evaluation form for student personnel services. In its various stages of refinement it has been used at a number of institutions with limited success. Recently, an "all-out effort" has been made to eliminate some of the remaining "bugs" and we feel that now we may have an instrument which is reasonably valid and which can be functional in the evaluation of student personnel services.

Actually, the form has been designed with the thought in mind that it might well be used in conjunction with an evaluation form which Dr. Wrenn and I described in an article in *School and Society* in 1948 (6). The earlier form, entitled "An Evaluation Report Form for Student Personnel Services" is for the use of trained personnel workers and combines judgments with regard to institutional philosophy toward the program and actual evidence of specific services. The present form, used in conjunction with the previous form, should give a comprehensive evaluation of a student personnel program, in that reactions of both students and the trained personnel worker are utilized.

Often judgments are made, relative to the value of a service, on the basis of a few students' reactions to a question or two. The present form is based upon the principle that *if several pertinent questions about a particular student personnel service are asked of a sufficiently large random sample of the local college population, a valid indication of the worth of the service to those students will be available.*

Sixty questions, five for each of twelve different services, comprise the present form. The twelve services listed below are those ordinarily included in any balanced program. All are self-explanatory with the exceptions possibly of "Adjustment of the Institutional Program to Student Needs" and "Guidance in Student Conduct." The former illustrates the point of view that no institution can have an effective student personnel

program unless the institution as a whole is functioning in the interests of the same student needs that the personnel services are designed to serve. The five items in this area provide an indication as to whether or not the total institutional emphasis is in this direction.

"Guidance in Student Conduct" is so stated in an attempt to place a particular emphasis on discipline. This emphasis is a counseling and learning emphasis in which students respond to items which indicate their sense of the justice of disciplinary procedures, and the extent to which discipline is a learning experience. If the policies relating to student conduct are consistent with the belief that each student who violates a regulation should be counseled and helped to learn from the experience, with punishment following only (1) when punitive action seems necessary for learning and (2) when necessary to restrict in the event no learning seems possible, then discipline can be a personnel service.

The five items in each area have been designed with the thoughts of achieving (1) the maximum coverage and (2) the best possible representation of the service, using a minimum number of questions. The items have been reviewed with the above in mind by various trained workers in the student personnel field.

The twelve services and a sample item for each follow:

Recruitment and Admissions

Do you feel that, previous to your admission, representatives of this institution adequately explained to you the facilities of this campus?

New Student Orientation

Do you think that this institution made you as a new student feel a part of it and of its activities?

Counseling Services

Do you feel that students on this campus who most need counseling are receiving such help?

Health Services

Are you satisfied that your campus health authorities would handle your case competently, in the event you were injured or became seriously ill?

Housing

Do you feel that this institution is making sufficient effort to improve student housing facilities?

Food Services

As a rule, do you feel satisfied with the food served you at the campus cafeteria or dining hall?

Extra-Class Activities

Do you feel that there are enough student organizations and activities on the campus to meet the needs of the students?

Adjustment of the Institutional Program to Student Needs

Do you feel that your total college or university experience is such as to better prepare you for intelligent citizenship?

Student Financial Aids and Part-Time Employment

If you were "financially on the rocks," would you feel free to go to the campus financial aid service for help and counsel?

Placement Services

Is your placement office making sufficient effort to keep you informed of current employment trends and needs?

Student Personnel Records

Are you of the belief that you are welcome to discuss with a counselor all matters contained in your student personnel folder?

Guidance in Student Conduct

Will a student on this campus get a chance to explain his side of the case if he is "called up" for discipline?

The sixty items as they appear in the form have been randomized in order to minimize bias and to insure a maximum chance that each item will be answered independently.

Administration of the Form.—It is recommended that a random sample of at least 200 students of the local college or university population be utilized in any study involving the use of this form. In order to determine the needs of various groups on campus, participants in the study are asked to check those of the following which are appropriate for them.

_____ Male	_____ Freshman	_____ Live Off-Campus in
_____ Female	_____ Sophomore	Rooming House
	_____ Upperclassman	Live at Home
	_____ Transfer Student	Live in College Dormi-
		tory
		Live in Fraternity or
		Sorority House
Major Department _____		

If one is to have a sufficiently large N from which to form judgments when considering any one of the above areas, it is necessary to have a reasonably large sample with which to begin.

Participants in the study may indicate "Yes," "No," or "?" in answer to each of the sixty questions. All items are so worded that if the service is functioning properly in the judgment of the student the "Yes" will be checked. If the service is inadequate, the "No" will be indicated.

The "?" is meant for use only in those cases where the student has insufficient knowledge of (or experience with) the service to make a "Yes" or "No" response. If an informed judgment of the adequacy of a service cannot be made, then use should be made of the "?".

Students are not asked to write their names on the form—only to answer the questions honestly and thoughtfully.

Scoring of the Form.—A Tally Sheet is provided which allows for (1) the tallying of responses to each item and (2) the grouping of these item responses for each of the services. (Each of the twelve services has a maximum "Yes" score of 500 for each one hundred students who participate in the study.)

Following completion of tallying, numbers of "Yes," "No," and "?" responses should be converted to percentages, using as the base N the total number of students participating. If one is considering only the responses of a sub-group, the number of students in that group should be used in computing the percentages.

If one wishes to consider only the "Yes" and "No" responses, i.e., only the *definite judgments* relative to the adequacy of the service, then one will need to use varying N's in computing the percentages, due to the probable variation in "?" responses for the twelve service areas.

Interpretation of Data: - One can assume that the higher the percentage of "Yes" responses for a particular service, the more adequate that service likely is in the judgments of the students. It is suggested that the services be regarded as adequate if the "Yes" responses approximate two-thirds or more of the total responses. (This is an arbitrary figure and a lower or higher percentage may be used if desired.) If less than two-thirds of the participants believe that the service is adequate, in terms of the five aspects of the service represented by the five items, then that service should be examined.

The "?" responses are to be used when there is a lack of familiarity with the service. The presence of even a low percentage of such (15 per cent or over, let us say) indicates the need for better lines of communication to the students. Often students are poorly informed as to the existence of the services that are provided for them.

The presence of a considerable number of "?" responses should not be interpreted to mean inadequacy of the service itself, but, rather, to be indicative of the need for a program of selling and of informing students of the services available. Actually, to have a strong program of student personnel services means little unless the various aspects of the program are known and are functional in terms of meeting student needs.

It is to be expected that underclassmen will indicate a lower percentage of "Yes" responses in the area "Extra-Class Activities" than will upperclassmen. Acquaintance with, and opportunity for participation in extra-class activities, generally increase the longer one is on campus.

Likewise, the percentage of "Yes" responses in the area "Placement Services" should be greater for upperclassmen. This service is especially designed for those approaching graduation and has less meaning for underclassmen. A high percentage of "?" responses should be expected of underclassmen in this area.

The evaluator may wish to compare the percentages of "Yes," "No," and "?" responses of one group on campus with those of another (for example, dormitory personnel with off-campus students). By utilizing appropriate tests of significance, one can be confident that differences found to be statistically

significant are real and not the result of chance errors of sampling. With such evidence at hand, one's conclusions will have greater meaning than they would, were there no statistical treatment of the data.

Finally, in the interpretation of data, it is well to keep in mind the goals and particular emphases of the institution. If, for example, the college or university provides a limited budget for a particular service or for the entire organized student personnel program, then it is probable that there will be a definite ceiling on the percentage of "Yes" responses for that service or program. Mention is made of the above because of possible criticism which may be inappropriately directed at certain capable student personnel workers who have inadequate programs due to insufficient institutional support. On the other hand, one must always be objective and critical of any low "Yes" response and examine carefully the service to see if the maximum is being achieved within the framework and limitations provided by the institution.

Summary

In order to ascertain the worth of a product it is well to question the consumer of the product. Such is true with regard to student personnel services. Accordingly, a student reaction form, containing sixty questions, five each for twelve commonly accepted student personnel services, has been devised. Through study of the proportions of favorable and unfavorable responses to the questions asked, one can determine certain program strengths and weaknesses, insofar as students are concerned. Use of the present form also permits one to secure data relative to the institution's success in actually making known to students the student personnel program it offers.

REFERENCES

1. Blaesser, W. W. "The College Administrator Evaluates Student Personnel Work." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, IX, Part II (1949), 412-428.
2. Brumbach, A. J. and Smith, L. C. "A Point Scale for Evaluating Personnel Work in Institutions of Higher Learning." *Religious Education*, XXVII (1932) 230-235.
3. Hopkins, L. B. "Personnel Procedure in Education." *Educational Record Supplement*, No. 3. Washington: American Council on Education, 1926.

4. Rohrer, J. H. "An Evaluation of College Personnel Work in Terms of Current Research on Interpersonal Relationships." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, IX, Part II (1949), 429-443.
5. Williamson, E. G. and Sarbin, T. R. *Student Personnel Work in the University of Minnesota*. Minneapolis: Burgess Publishing Co., 1949.
6. Wrenn, C. G. and Kamin, R. B. "A Procedure for Evaluating a Student Personnel Program." *School and Society*, LXVII (1948), 266-269.

THE MEASUREMENT OF STUDENT CONCEPTIONS OF THE ROLE OF A COLLEGE ADVISORY SYSTEM

EDGAR Z. FRIEDENBERG

University of Chicago

MOST colleges and universities provide some kind of counseling service for students. These services appear to have stemmed primarily from two functions: the organization of a student's program in such a way that requirements for degrees and for admission to professional schools may be met efficiently, and the enforcement of regulations deemed necessary by the college for the discharge of its responsibilities to students, parents and community. In many schools little connection has been perceived between these functions; program-planning occurs at registration, under the direction of the Faculty; breaches of regulations or aberrant and unsocial student behaviours are treated as disciplinary problems by the Dean of Students, or, even, separately by sexes in the office of the Dean of Men or Women, as in each case is appropriate.

With the increased influence of psychology on professional education (3) has come greater insight into the unity of the educable personality (4, 5). As a consequence, the division between emotional, disciplinary, and academic problems has been perceived as unreal (1, 8). Students make vocational choices based on fantasy or emotional tension; students fail in programs because of intrapunitive personality trends, hostility to authority, or inferiority feelings; students behave lawlessly out of an appetite for punishment which grows on what it feeds on, or in acting out fantasies so complex and deep-seated as to render disciplinary action, however severe, an extraneous factor whose meaning is distorted by the same mechanism which precipitated the behaviour. Not all students do these things, of course, but, unless the admissions policy of a college is so ineffective or rudimentary as to admit large numbers of students

who are simply too stupid to succeed, or too poor to have time to study after they finish their part-time work, it is clear that emotional factors must be involved in most of the academic or disciplinary problems which do occur, whether these are confined to a small group of students or are prevalent in the student body as a whole.

Even so, however, there remains the fundamental question of the degree of responsibility which a college has for the emotional welfare and personality structure of its students, and the administrative question of how such responsibility is to be discharged, if accepted. It is always possible to set up quasi-clerical bodies whose function is to excrete unsuccessful or unconforming students. Education is, however, a systematic process by which human behaviour is changed in directions which the student accepts and the faculty deems good and desirable. To limit the techniques of changing behaviour to those which can be applied from the lecture platform, and the effectiveness of education therefore to those who can immediately, realistically, and maximally profit by those techniques, seems short-sighted and intransigent and, in many cases, cruel. To do so with a student body composed in large part of youngsters seems irresponsible.

At the University of Chicago, whose College, as is well known, accepts students after the second year of conventional high school, no such limitation has ever been considered. There is a complete student health service, extending from orthopedics to psychiatry. There is a Counseling Center, using client-centered techniques (6), to which any student can turn, without charge, for assistance in "thinking through" questions with which he is concerned. There are conventional vocational guidance services. There is not, since all University facilities are thought of as contributing ultimately to intellectual development, a University Mortician; one can only say in defense of the lacuna that few students develop a need for the services of such an official while in residence and none has ever applied for them.

Within the College of the University, and peculiar to it, is also the College Advisory System. This consists of a staff of approximately (the number varies slightly from year to year)

twenty advisers in the College, usually devoting from one-fourth to one-third of a full-time assignment to the advisory service, and carrying a case load averaging 120 students. While it cannot be said that a systematic philosophy of advising underlies the system, it has adhered to certain principles since its inception. One of these is that all advisers shall be primarily members of the College Faculty, devoting their major effort to teaching or research. The purpose of this is to insure familiarity with the operations of the College, so that the adviser may discharge his administrative functions accurately. Another is that students need not be assigned to an adviser of the same sex, the purpose of this policy being to dispel the atmosphere of obsession with the erotic which has characterized many student personnel services of a more conservative orientation. A third is to assign each student, so far as possible, to an adviser with special qualifications in his intended field of professional or academic specialization; but since students in the College have virtually no opportunity to modify their programs of general education so as to contribute directly to their vocational goals, this policy has been modified so that students are not assigned special advisers until they are near to the completion of their work in the College or have made definite plans for advanced study or professional training. Students are not allowed to choose their adviser, but, are usually, on their request, removed from the list of the adviser to whom they have been assigned and placed on the list of the adviser whom they prefer, or whose special academic field is the one in which they are most interested, if he has room for them.

New students normally meet their advisers for the first time at a twenty-minute registration conference at the opening of the year; students are admitted to the College only at the opening of the Autumn Quarter. At this time the adviser registers the student for an entire year, and files, without discretion, on the basis of placement-test results, the student's program for the Bachelor's degree. No administrative device has succeeded, despite much thought and worry, in making these conferences anything but rushed and unsatisfactory; an inept adviser can, during registration, infuriate, confuse, or frighten as many as sixty new students, although the average

is doubtless somewhat below this number. After registration, students may sign up for a fifteen-minute appointment with their adviser at any time they wish, for any reason they wish, during the period of eight to ten hours per week which the adviser allots for the purpose. They may also be summoned to see the adviser at his discretion, almost always to discuss academic problems. The adviser's signature must be obtained to any change of registration initiated by the student.

It may be seen, therefore, that the College Advisory System operates in an almost totally academic context. In a school which admits only intellectually qualified students, and provides fairly generously for assistance to those who need it, most academic problems, however, seem to originate in a disordered perception by the student of his situation and responsibilities, accompanied, of course, by the underlying anxieties and regressions which give rise to the need to misunderstand. There is a question, then, as to how much insight into the emotional origins and significance of academic problems a subject-matter specialist can be expected to acquire in order to be most helpful in solving them. But there is a deeper and more controversial question than this on which the responsible adviser must take a position. In every college there are a number of students whose academic success is enhanced, rather than hindered, by aspects of their personality which seem likely to result in great ultimate unhappiness. There are students who use preoccupation with abstract theoretical material to distract themselves from personal and social inadequacies. There are students who seek academic distinction in order to flaunt it in defiance of a culture which they believe to disparage it. There are students who are convinced that they can only be valued because of their scholastic achievements, and who are ceaselessly driven to seek grades as copper tokens to exchange for affection at a very unfavorable rate. What is the responsibility of the adviser for the welfare of such students? Must he train himself to recognize them? If he can recognize them, should he seek to initiate personality changes which will probably make the student's academic record less spectacular, even if they also result in greater happiness and ultimately greater productivity and creativeness?

The answer to such questions depends on a complex hierarchy of values, which certainly cannot be established by empirical investigation alone. It is clear, however, that student expectations of the Advisory System are one of the factors which must affect the decision. No administrator can build an advisory service in response to student demand, which is always partially conflicting and made in partial ignorance of the administrative limitations of the particular situation. If, however, a certain kind of service is believed by students to be a responsibility of the Advisory System, although no administrative provision is made for it, a situation which will engender hostility, and which is dangerous if the service is important, exists. On the other hand, if students are convinced that a particular kind of service is *not* the responsibility of the Advisory System, and would not seek it there even if it were offered, that service can probably not be offered to students effectively within the System, particularly if it is a counseling service which must, ultimately, always be voluntarily received.

The author, therefore, sought to develop an instrument which would measure four things: (1) student opinion of the *scope* desirable in the College Advisory System; (2) student information about the system as it actually exists, to permit an estimate of the degree to which criticism and opinion might be regarded as informed; (3) student evaluation of the effectiveness of the System in solving certain problems which it recognized as possible sources of weakness in itself; and (4) an indication of the kind of *role* with respect to themselves students believe an adviser should play in assisting in the solution of certain complex problems. Since this information seems to be among that which would be needed by any college in evaluating its advisory services, the instrument used to gather it will be described and illustrated in some detail. (Copies of the complete instrument may be obtained from the author on request.) It consists of a group of five batteries of objective questions, with space provided for additional focussed written comment by students; the entire instrument requires something under two hours for most students to complete. The first battery consists of nine questions which elicit only vital statistics—age, position in the college, frequency with which student consults adviser, etc.

Because of the unique mode of organization of the College, few of these questions would be applicable intact to other situations, and they will not be reproduced here. Since IBM electrographic answer sheets were used, questions were numbered so as to facilitate analysis, and the next battery began with item 16. Most of it is reproduced, as follows:

Below you will find listed certain problem situations which are encountered with varying degrees of frequency among College students. Among the resources to which a student at the U. of C. might turn for assistance with each of these problem situations is his College Adviser. In considering each problem situation, feel free to draw on your own experiences with the College Advisory System, or other information which you believe to be valid, but try in every case to give a reasonably *generalized* response, based on your conception of the system as a *whole*. For each of the situations listed, on your answer sheet *blacken* space

- A. if you believe the College Adviser to be the *best* person from whom to seek help in such a situation.
- B. if you believe that the College Adviser would be the *best University staff member* from whom to seek help in such a situation, though probably less effective than experts available elsewhere (e.g., a private psychoanalyst or firm specializing in vocational placement).
- C. if you believe that the College Adviser might be of *some help* in such a situation, and that you might go to him if you had special respect or friendship for him, but believe that there are other more appropriately trained and chosen University officials who could be of greater assistance.
- D. if you believe that *some* University official should be available to help in such a situation, but that a *College Adviser*, either because of deficiencies in training, insight, or interest, or because his responsibilities are divided between the student and the institution, *might be an indifferent or even dangerous source* from which to seek it.
- E. if you *cannot conceive* that the University has any responsibility to help a student with such a problem, and do *not* believe that this student should seek help from *any* University official.

PROBLEM SITUATIONS

- 16. Student is fearful of failing his comprehensive examinations, even though he has been working and has made passing grades in the Autumn and Winter Quarters.
- 17. Student must work to remain in school, and finds that in order to clear enough time to keep a job, he must petition to get into sections of classes that are listed as closed.

18. Student has stolen an automobile and later abandoned it. He has not been detected, but fears that he may be, and anxiety is disrupting his work and his life.
19. Student is making mostly *C*'s, with an occasional *D* and still less frequent *B*. The death of his father makes it impossible for him to continue in school without substantial financial aid.
20. Student cannot bring himself to study; if he sits at his desk and attempts to do so, his mind wanders off into day-dreams. If he attempts to write a required paper, or other written exercise, the blocking is particularly intense.
21. Student wishes to enter medical school in the shortest possible time, and wants help in planning his program of studies most efficiently.
22. Student is uncertain whether the qualifying examination in Humanities 1 (Special Art) can be taken as part of a sequence culminating in Humanities 3 (German) in fulfillment of the requirements for the A.B. degree, and if so, whether Language 1 is still a requirement or not.
23. Student has gotten into serious difficulty as a consequence of sexual relations, and is now in a state of panic at the prospect of having to choose between an undesired marriage or exposure and parental discipline.
24. Student, not living in a residence hall, has participated in a group which went to a Gerald L. K. Smith meeting to break it up. Eggs were thrown, and the student is now being held by the police.
25. Student has a mild interest in becoming a lawyer, which is in accord with his parents' wishes. He is not certain that his interest is very real, or that he has the pattern of abilities which lead to success in this field, and is beginning to feel anxious.
26. Student is troubled with severe headaches, of undetermined origin, which are making it impossible for him to study and causing him to fail his work. He notices that they are followed by periods of listlessness and depression.
27. Student has purchased a portable typewriter from a store in the University community, and signed an installment contract to pay for it. He has found several mechanical defects in the machine, and wishes to return it and get his money back. The store, however, threatens to sue him for the balance of the money.
28. Student does not understand the process by which his placement has been made and wishes to have the meaning of his placement scores explained to him, as he feels he should have been excused from Mathematics 1 and Social Sciences 2.
29. Student has developed a very strong emotional attachment to his roommate, who is now no longer willing to "pal around" with him as he did at first. The roommate has

requested a change of room assignment, and the student is troubled by suicidal impulses, and terrifying dreams in which he is murdered by his former friend.

The reader will doubtless grant that nearly every type of problem is represented in the battery, from the purely academic to the highly aberrant and clinical. These last were included, not because an adviser is likely to encounter them but in order to permit students to express the most extreme demands possible on an Advisory System if they wished.

The next battery the only portion of the instrument to which a right answer "key" in the usual sense of examining is possible—consisted of twenty true-false statements about the Advisory System. Examples are "Penalties may be invoked to compel a student to register for those College courses which his adviser recommends that he take during a particular year," "Most College advisers carry a case load of approximately fifty students," "College advisers receive special training in the realistic handling of the emotional problems of students."

The fourth battery, consisting of 15 questions, would be adaptable to almost any academic situation, and is reproduced below in its entirety.

In the College Advisory System, as in every administrative structure, the performance of the functions characteristic of that system is limited by problems of facilities and procedures. Sometimes these limitations can be overcome by ingenuity and special technique; often they persist as sources of dissatisfaction to staff and clients alike.

Below you will find listed a series of such limitations which you may or may not feel apply to the College Advisory System. In considering each limitation, feel free to draw on your own experience with the College Advisory System, or other information which you believe to be valid but try in every case to give a reasonably *generalized* response, based on your conception of the system as a *whole*. For each of these, on your answer sheet *blacken* space

- A. if you feel that this problem is *almost always satisfactorily* overcome by the College Advisory System, or is *one with which it should not be concerned anyway*.
- B. if you feel that the problem is *often* satisfactorily overcome by the College Advisory System, but is nevertheless the source of *occasional* annoyance.
- C. if you feel that the problem is *recognized* by the College Advisory System, but is *mishandled* about as often as it is solved, or has been solved by *halfway measures*.

- D. if you feel that the problem is one which may *usually* be expected in contacts with the College Advisory System, although you are *occasionally* surprised by successful handling of it.
 - E. if the problem is *almost always troublesome* in student contacts with the College Advisory System to which it is related, and there is *no satisfactory* evidence of effective attempts to solve it.
61. Providing of enough time at each interview to permit students to complete the business for which they sought an appointment.
 62. Keeping individual advisers close enough to their schedules that students need not wait too long for their appointment, or miss class time because of late advisers.
 63. Finding persons to serve as advisers who are warmly interested in students and their problems, and who know their students as individuals.
 64. Keeping the case load per adviser low enough to permit advisers to get really acquainted with their advisees and their problems.
 65. Keeping student conference material confidential, and not revealing it to persons who might use it in damaging ways.
 66. Knowing accurately the right members of the University to whom to refer students with special problems—e.g., reading deficiencies, or presumed errors in recording comprehensive results—and helping students to get in touch with those people.
 67. Providing office facilities which insure as much privacy as students need in order to discuss freely with their adviser such problems as they wish.
 68. Assigning as advisers persons with sufficient insight into the emotional and developmental tasks of young people to really understand what's going on inside them.
 69. Keeping records sufficiently up-to-date, accurate, and available that advisers do not act on mis-information.
 70. Conveying to students an attitude of respect for them as people, and conducting interviews with courtesy and genuine friendly feeling.
 71. Getting advisers to shut up long enough to permit students to express their own feeling about problems fully.
 72. Assigning as advisers persons of sufficient maturity that they need not "use" students emotionally, by bullying, identifying too much with them and their problems, making demands on the student for liking or admiration, or in other, more subtle, ways.
 73. Providing advisers sufficiently mature emotionally to listen to any problem students might wish to discuss with them without becoming "shocked" or frightened, or attempting to impose standards of conduct which the student does not accept.

74. Scheduling sufficient hours per adviser that students can get to see an adviser when they need to, without having to wait for attention with their problem unsolved.
75. Providing sufficient information on "summons" forms that students are not caused needless anxiety as to the possibility that they may be in trouble.
76. Limiting the scope of the adviser's activity sufficiently that students are not obliged to discuss with him matters which are not properly his business.

The fifth battery, although it contains but five items, is perhaps the most interesting in the questionnaire. It is intended to appraise the *role* which students think it appropriate for the adviser to fill, and consists of fictitious case studies, each of which presents a rather serious student problem, followed by a choice of five courses of action which the adviser, confronted by such a problem, might take. The student is asked to indicate the choice he believes best, and is given space for written comments in which to suggest other courses he might judge preferable. The items follow:

91. Student is afraid that he will fail comprehensive examinations in German and Mathematics. In the course of his first interview with the adviser, he reproaches himself severely for his failure to study, but states that, as soon as he begins to try to do so, his mind wanders off into day-dreams. He is a good jazz musician, and is in demand by many of his former high-school friends to lead a small orchestra at their social events. When he agrees to do this, his parents attack him, pointing out that he has never been as smart as his elder brother, that he is wasting his time and their money, would probably have a hard time succeeding at the University of Chicago in any case, and must surely transfer to an easier school if he fails an examination.

The boy, as he tells this story, seems much hurt and uncertain, but is inclined to agree with the low estimate placed by his parents on his character and intelligence. Entrance aptitude test scores secured by the University place him well among the upper tenth of applicants admitted.

A good adviser would

- A. sympathetically but firmly support the parents' demands on the boy, advising him to give up the orchestra until he is more certain that he can carry his schoolwork.

- B. tell the boy unemotionally that the decisions must be his, but reiterate for him the precise requirements for continuing registration in the College.
 - C. say only enough to make it clear to the boy that his feelings of anxiety, rejection and conflict are understood and accepted.
 - D. sympathetically point out that the boy has a right to make any decisions about his total program of activities which will best satisfy him, while making sure that he understands both the conditions under which he may continue in school and the real abilities he has been shown to possess.
 - E. point out that the key to the situation is probably the hostility his parents feel toward him, as shown by their desire to underrate him, and his resultant fear that, should he succeed, they will completely reject him.
92. Student, an eleventh-grade entrant, seventeen years old, has been placed on probation because of a failure to attend required physical education classes. She is also failing two of her subjects. The instructor in one of these has turned in a sympathetic report, indicating that he believes the girl to be intelligent and creative, but too much burdened by her personality difficulties to accomplish much at this time. The other report is aggressively critical, describing the girl as unkempt and lazy, and declaring that she has no place in the College. At the conference to which she is summoned, the girl appears shy, nervous, and so far as possible, uncommunicative.

A good adviser would

- A. point out to her in a kindly but resolute way that she will surely be dropped from school if she does not make a better academic adjustment, and help her to schedule her week's work so that she can begin to make effective use of her time.
- B. restate to her, in as neutral a tone as possible, the conditions under which her registration may be terminated, but emphasize that the decision must be hers.
- C. let her know that he understood that she must be feeling threatened and unhappy and express clearly a wish to help her understand her own feelings better, while pointing out calmly that they must also meet the practical situation in which she is involved in order to go on working together.
- D. suggest that she drop the course taught by the hostile instructor, and use the extra time to catch up on her other work.
- E. point out to her that her unkemptness, laziness, and

uncooperative attitude are quite evidently ways of rebelling against authority and are almost certainly derived from her feelings about her parents rather than from any real aspects of her College situation.

93. The program of an 11th-grade entrant has been erroneously prepared by his registration adviser, who checked Biological and Physical Sciences rather than Natural Sciences 1, 2, and 3, as requirements for his degree. The error is noted shortly before the beginning of the student's second year in the College, and the student is notified that the requirement has been changed and that he must now take the Natural Sciences sequence. The student has not yet registered for either Biological Sciences or Physical Sciences, and could not have begun work on Natural Sciences 1 during the previous year because of poor mathematics placement, so that he has not, in fact, suffered as yet by the error. He is nevertheless quite upset by the change, as he wishes to enter an engineering school, believes that Physical Science will serve him in better stead than Natural Sciences 1, does not want to take an additional comprehensive, and is angry about the inefficiency of the adviser in making such an error. He comes in to ask that the original statement of his degree requirements be kept in force.

A good adviser would

- A. apologize for his carelessness in making the error, but point out that since it has not as yet affected the student's program, the requirement should stand as corrected.
 - B. state firmly that error or no error, the degree requirements for 11th-grade entrants are uniform and must be consistently administered.
 - C. note carefully the student's reasons for wanting to keep the old requirements in force, then take the matter to the Dean of Students in the College, admit that the original error was his, and ask the Dean to stand behind the old requirements.
 - D. himself prepare an amended program for the student, reaffirming the original requirement, and send a copy of it to the Registrar for recording.
 - E. point out to the student that it is irrational for him to be angry over an error which has, in fact, done him no harm, and try to help him to gain insight into the true sources of his annoyance.
94. An 11th-grade entrant has a schedule which requires that he take Physical Education at 1:30. He schedules a conference with his adviser at which he complains, with some indignation, that this program is not acceptable to him, because it interferes with his freedom of worship. It has

been his custom, since the age of ten, to read a chapter of a religious work daily after lunch; if he does not do so, his food disagrees with him, and he suffers from bloating and heartburn. He believes it to be dangerous to his health to take exercise while in this condition, but maintains stoutly, and unasked, that this does not bother him at all, since he is prepared to meet his Maker at any time. He does, however, insist that, rather than risk the moral obloquy thus involved, he will simply refuse to attend physical education classes. There is no way to arrange his schedule so that he can either lunch at 11:30 or take Physical Education then without either petitioning for admission to three closed class sections or getting the Physical Education Department to make an exception to its rule and let the student come two days a week at 11:30 and two days at 1:30.

A good adviser would

- A. let the boy go ahead and petition, regardless of the improbability that three petitions would be granted for such a reason, in the hope that he might change his mind when finally confronted with so nearly impersonal a reality.
 - B. attempt to persuade the Physical Education Department that the boy's emotional need is important and real, and that it should make an exception in this case.
 - C. say neutrally and dispassionately to the boy that the University does not recognize this kind of fantasy as religious in character, and cannot accommodate itself to such diversity of need; tell him frankly that if he does not attend compulsory physical education classes, he will be removed from the College.
 - D. tell the student that it is pretty clear that some factor besides religious conviction is operating to produce symptoms of this kind, that the responsibility of the University to him and his parents requires that it insist he report to Student Health for a complete medical and psychiatric examination, and that his program may more profitably be discussed in the light of the report which Student Health will make.
 - E. discuss with the student the religious meaning of his position, pointing out that it must derive from an unusual conception of God, and suggesting that he scrutinize his own emotional needs as the source of the conflict.
95. A twenty-year-old student, who entered the College at the 13th-grade level at the opening of the previous scholastic year, is making satisfactory grades, both on his comprehensives at the end of his first year and on quarterly ex-

aminations. Reports from his instructor in Humanities 2 and History of Western Civilization commend him for his brilliant contribution to discussion, and his evident capacity to integrate the material offered into abstract generalizations. Reports from his instructors in Biological and Physical Sciences indicate that he has hardly ever attended classes in these courses, although he passed the comprehensive in Biological Sciences with a grade of C.

The student's adviser, in an informal discussion with the Head of the residence hall in which the student lives, learns, however, that the student is regarded by the Head as somewhat lacking in emotional adjustment. He has taken no interest in House social activities, and, so far as is known, has few social interests of his own. His friendships within the House are confined to two other boys, with whom he has discussions nearly every night centering on the Marxist interpretation of the motivations of contemporary politicians, or the unity and structure of contemporary drama, or the nature of reality. He has twice been sent back to his room from the dining hall because he came in to dinner without coat or tie.

A good adviser would

- A. do nothing about the situation, on the grounds that he has no right to interfere with what evidently represents the boy's free choice of behavior, so long as he is academically successful.
- B. summon the boy for a general discussion in the course of which he would expect to describe to the boy in detail the range of interesting activities available at the University.
- C. attempt to show the House Head that the behavior of the boy might very well indicate more complete achievement of the objectives of the College than that shown by nominally better adjusted students, and urge him to encourage the boy's present mode of self-expression.
- D. summon the boy for a conference in which he would cautiously attempt to estimate how happy the boy really was, and, if considerable anxiety and unhappiness were indicated, try to get him to discuss the possibility of seeking help from the Counseling Center or a psychiatrist.
- E. summon the boy and explain to him that his present behavior shows serious maladjustment, is probably more the result of his need to rebel against the patterns of middle-class behavior established by his parents than of a serious interest in his studies, and suggest that he work the problem through with the adviser.

Detailed results of the administration of the instrument will not be presented here, since it is hard to see how they would be of more than local interest. A brief account will be given, however, as an example of the way the questionnaire may be handled, and the kind of results to be expected from it.

A letter describing the questionnaire, and stating that it had been prepared jointly by the Offices of the Dean of Students and University Examiner was sent to every seventh student on the list of each adviser, requesting him to come fill out the instrument at his choice of four specified times. Since independent results were wanted, this seemed more desirable than sending the instrument to the student, who would, in many cases, have then filled it out in consultation with others. At the time this was done, the *Chicago Maroon*, the official student newspaper, editors of which had been present at all sessions where the questionnaire had been planned, carried editorials urging student co-operation. 161 students or slightly less than half of those who were invited, filled out the questionnaire. The composition of this sample was scrutinized by the Dean of Students in the College, who declared it to be adequately representative, so far as crude statistical factors, i.e., length of residence in the college, age, sex, level of admission, etc., were concerned. The sample could *not*, however, have been representative of student attitude, since it is quite clear that the large proportion of students who did not respond must have felt differently about the Advisory System than those who were willing to give it some time. One would assume, in the absence of more specific information, that students who felt most strongly about the system, whether positively or negatively, would be likely to respond, while the indifferent would ignore the request; such an inference could be checked only by an aggressive interviewing program in which contact was established with a good sample of those who refused to co-operate.

The results on the objective portion of the instrument cited in this article, for the total group of 104 boys and 57 girls responding, are presented in the following table. For items 16-29, the figures given refer to the number and percentage of students marking the item A, B, C, D, or E. For items 31-50, the figures are a frequency distribution showing the number of

students making various total scores on this twenty-item true-false "test" of information. For items 61-75 the same information is given as for 16-29, with two additions. These items, as the reader may perceive by referring to them, constitute a rating scale on which students appraise various problems which the Advisory System may have met more or less effectively. Space A represents a highly favorable appraisal on a particular item, space B a moderately favorable one, space C a neutral or ambivalent one, space D moderately unfavorable, and space E highly unfavorable. In order to provide some quantitative indication of the relative success of the system in solving these problems, the following device was invented. The number of students choosing to rate each item A was multiplied by 3; the number of students marking it E, by -3. Those marking it B were counted in as 1, those marking it D, as -1, while C responses were ignored - multiplied by zero. The total sum thus obtained was added algebraically for each item, and the number thus obtained is reported as a Derived Score in the column D.S. The Rank column simply indicates the rank of these scores, a low number indicating a highly favorable student response to this aspect of the service. The maximum possible score would be 3×161 , or 483; the minimum -483. Astonishingly, but gratifyingly, no negative scores are obtained.

Similar data have been gathered for six subgroups of the population which took the questionnaire. These groups are: 54 1948 entrants, who had had but a few weeks experience with the College; 41 11th- and 12th-grade entrants aged 18 or younger; 31 students having been assigned to three or more advisers during the course of their college career; 62 students answering correctly eleven or fewer of the twenty true-false information items; 35 students blackening space E (maximally unfavorable) for two or more of items 61-76; and 97 students choosing unpopular responses - that is, responses other than 91D, 92C, 93A or C, 94B or D, or 95D on two or more of the "case-study" items 91-95.

Three different kinds of free responses were sought from each student. The first and major source of these was the following

TABLE 1
*Performance of 161 Students Responding to Invitation to Complete the Questionnaire
Evaluating Their Conception of the College Advisory System*

Item	Responses									
	N	A %	N	B %	N	C %	N	D %	N	E %
16	57	35.4	30	18.6	58	36.0	10	6.2	5	3.1
17	134	83.2	4	2.5	15	9.3	6	3.7	0	0.0
18	3	1.9	28	17.4	30	18.6	47	29.2	51	31.7
19	60	37.3	11	6.8	60	37.3	19	11.8	9	5.6
20	9	5.6	74	46.0	35	21.7	34	21.1	9	5.6
21	106	65.8	2	1.2	40	24.8	12	7.4	0	0.0
22	153	95.0	1	0.6	5	3.1	0	0.0	0	0.0
23	1	0.6	17	10.6	29	18.0	43	26.7	69	42.8
24	5	3.1	8	5.0	53	20.5	39	24.2	72	44.7
25	24	14.9	46	28.6	71	44.1	15	9.3	2	1.2
26	0	0.0	54	33.5	45	28.0	43	26.7	17	10.6
27	12	7.4	10	6.2	44	27.3	36	22.4	56	34.8
28	143	88.8	2	1.2	12	7.4	3	1.9	0	0.0
29	0	0.0	66	41.0	33	20.5	46	28.6	16	9.9

Distribution of Scores—Items 31-50

Score	N	%
0-1	0	0.0
2-3	1	0.6
4-5	2	1.2
6-7	8	5.0
8-9	11	6.8
10-11	40	24.8
12-13	43	26.7
14-15	44	27.3
16-17	10	6.2
18-19	1	0.6
Omit	1	0.6

$$M = 12.2 \sigma = 5.04$$

Item	Responses										Rank	D.S.
	N	A %	N	B %	N	C %	N	D %	N	E %		
61	65	40.4	68	42.2	16	9.9	5	3.1	6	3.7	10	204
62	47	29.2	83	51.6	9	5.6	14	8.7	6	3.7	9	206
63	24	14.9	60	37.3	42	26.1	21	13.0	10	6.2	13	81
64	21	13.0	44	27.3	46	28.6	58	17.4	17	10.6	15	28
65	120	74.5	24	14.9	5	3.1	4	2.5	0	0.0	2	380
66	67	41.6	50	31.0	26	16.1	8	5.0	4	2.5	8	231
67	64	39.8	41	25.5	25	15.5	7	4.3	18	11.2	11	172
68	26	16.1	45	28.0	40	24.8	29	18.0	14	8.7	14	52
69	67	41.6	66	41.0	13	8.1	9	5.6	1	0.6	7	255
70	98	60.9	45	28.0	12	7.4	4	2.5	0	0.0	3	335
71	46	59.6	49	30.4	5	3.1	5	3.1	4	1.2	4	326
72	115	71.4	37	23.0	4	2.5	1	0.6	0	0.0	1	381
73	91	56.5	40	24.8	12	7.4	8	5.0	2	1.2	6	299
74	27	16.8	64	39.8	34	21.1	14	8.7	20	12.4	12	85
75	31	19.2	43	26.7	25	15.5	14	8.7	37	23.0	16	21
76	101	62.7	24	14.9	9	5.6	2	1.2	3	1.9	5	318
91	10	6.2	5	3.1	0	0.0	135	83.8	5	3.1		
92	22	13.7	5	3.1	110	68.3	12	7.4	3	1.9		
93	54	33.5	1	0.6	67	41.6	5	3.1	24	14.9		
94	10	6.2	49	30.4	1	0.6	71	44.1	18	11.2		
95	33	20.5	23	14.3	10	6.2	80	49.7	1	0.6		

paragraph, presented at the close of the objective portion of the material

On this sheet please suggest any specific changes in the College Advisory System which you believe would increase its effectiveness. Feel free to suggest any that seem important to you. It is suggested that you center your thinking around such possible areas of change as:

1. Professional qualifications of advisers.
2. Case load of advisers.
3. Scope of advisory service. i.e., increasing or decreasing the range of *kinds* of problems with which advisers deal. Do you feel that advisers, as they now function, are a threat to freedom or privacy of students? Do you, on the other hand, feel that they are too much concerned with routine academic problems to offer you the help you need? What changes would you suggest?
4. Intercommunications between Instructors, House Heads, and Advisers.
5. Means of establishing the working relationship between student and adviser as soon as possible.

Students were also asked to list any characteristics of the Advisory System not included in items 61-76 which seemed to them especially worthy of favorable or unfavorable comment, and to state any specific course of action which they would prefer to any of the 5 listed, with reference to items 91-95. These comments have been examined rather carefully, and, so far as they are subject to classification, tallied quantitatively.

Twenty-five (of the 161) students responding to the paragraph quoted above expressed a feeling that the case load of advisers should be limited—the most common single suggestion made. Twenty-three felt that advisers should be warmly interested in students, and 17 felt that more attention should be given to personal problems of students. Sixteen students felt advisers should receive training in adjustment counseling procedures, or psychology, while five more also felt this to be the case if emotional problems of students were really a part of the advisory responsibility, but were not quite prepared to concede that they were.

On the contrary, a smaller group of students displayed contrary and apparently more intense feeling. Eleven students felt that advisory and counseling services should be kept separate, or that the adviser should not be concerned with personal prob-

lems, while one student put the feeling on the basis that advisers should not deal with problems which students would ordinarily discuss with parents. Fifteen students took what might be termed a middle position, viewing the advisory function as mainly academic, but feeling that advisers should be able to direct students intelligently for help when needed. A similar and related contrast was apparent in student wishes concerning the degree of interrelationship between advisers and dormitory and academic staff. Fourteen felt this should be increased, while nine felt this to be undesirable.

Of particular interest was the extent to which students conceived the Advisory System as playing an important role in interpreting the purposes and values of the University of Chicago College Plan to them—a function which, it must be admitted, was almost completely ignored in the instrument itself. This feeling was expressed in a variety of ways, and is therefore less conspicuous on the tally than it would have been had the attitude found expression in a single, often-repeated sentiment. Nine students expressed a direct wish for assistance in the synthesis and interpretation of their College learning experiences. Six, expressing less positive feeling, expressed a need for more assistance in orienting themselves to the University. Six also wanted this help specifically in connection with the function of the Advisory System itself, with two expressing definitely the feeling that the System should more clearly define and state its own purposes and limits. Some anxiety was expressed at the failure of the College to take students' vocational ambitions adequately into account, five felt that special advisers trained in a professional field, e.g., for pre-medical students, should be assigned as needed, four, that more help should be given the student in making plans to enter a Division on completion of general education. There was surprisingly little complaint about the non-voluntary nature of the system, only four students wished to be allowed to choose their own adviser, while nine asked that regular meetings be scheduled at intervals, regardless of their felt need, in order to check on their progress.

Few items were added to those in 61-76 of the instrument by student commentators, and none by more than four persons

Rather curiously, four students stated here a belief that students should have an adviser of their own sex; three wanted more technical advice about planning for a vocation, three felt that advisers should be commended for trying to help students with personal problems, and should do more of it; and three felt that warmer, friendlier relationships would be desirable, two wanted more psychological or clinical training. On the other hand, three students felt that advisers should stick to impersonal or academic problems, and not pry into others, and two, that academic and emotional counseling should be kept separate.

Reactions on the case study items were, as was expected, interesting and revealing. Seventeen students, as might be expected, recommended referring the boy of item 91 to a source of more specialized psychological care, in most cases medical. Fourteen, however, wished the adviser to intercede directly with the parents to get them to understand the boy better. Other recommendations were largely partial or palliative; financial aid, so that he could live away from home, assistance in scheduling, and the like. Four specifically enjoined the adviser to follow through on what would evidently be a long and difficult case.

On item 92, psychiatric aid, recommended by 27 students, was virtually the only cogent suggestion to emerge. Three students, however, recommended a stern attitude.

Perhaps the most striking characteristic of the responses on item 93 was their almost uniform hostility. Only five students recommended that the adviser attempt to get the consent of the University to the maintenance of the existing, erroneous agreement as the student wished, which was, indeed, the course of action successfully undertaken in a closely parallel case which suggested this item. Nine students urged that the adviser politely but firmly require the student to take the Natural Sciences program. Fourteen recommended that the adviser explain to the student the advantages of the Natural Sciences sequence, and its greater consonance with the objectives of the College. Three students recommended an aggressive firmness—one of these stating that "a few spankings when he was younger" might have helped the student, and another stating that he

should "know when to keep his mouth shut " As the item, as presented, gives no intimation of the personality of the student involved—intentionally so, since this item was chosen to measure student reaction to a purely administrative situation without clinical aspects—this evidence suggests that many students in the College are highly identified with its objectives, and highly intellectual character, (7) and are inclined to exempt the College Plan in the abstract, though not the staff or administration, from duty as a target for rebellion

Great, indeed, is the contrast presented by responses to item 94 While eight students recommend referring the boy to a psychiatrist, and three suggest that he be required to conform, seven state that the program must be changed, because not to do so would infringe on the boy's freedom of religion; five, in this case, *suggest direct appeal to the administration to insure that case is not handled legalistically* One student states that the adviser "must do everything possible to maintain the boy's faith " Three suggest that the boy be referred to an official of his own church

On item 95, as might be expected, most of the commentators, as would be expected, were concerned about the possibility that the student might be coerced, or that his privacy might be unduly invaded, than were concerned about his ultimate fate In the main, they were not hysterically so, and there was considerable acceptance of the dangers which such a student might be piling up for himself Nine students felt that interference of any kind was unjustified, or that the behavior of the boy was not peculiar Five, however, thought counseling should be given; three, that the student should be introduced around, and six, that his old interests should not be discouraged, but that he should be led to develop new ones It should be emphasized with reference to this item, as with the other four in its group, that students did not make comments unless they wished to amplify or reject the five already available to them in the instrument, reference to Table 1 will show that most of the students were able to accept one of the positions presented in the item

What inferences do these data suggest, with reference to the questions raised as to the scope and responsibilities of a col-

lege advisory system? Perhaps the most interesting and suggestive is the rational picture which students seem to have of the Advisory System and of its limitations. They recognize that, in a situation providing as varied services as the University of Chicago, its function is primarily academic. This would not seem to indicate, of course, that students regard the degree of insight into the sources of their difficulties which an adviser can muster as unimportant, rather, that they do not expect advisers to develop sustained clinical relationships with them. Evidence for this comes both from responses to items 16-29 and from the "case study" items.

Nevertheless, many students who are well aware that certain problems are psychiatric, and that advisers are not psychiatrists, still consider that the University has a responsibility to assist them with such problems, and believe the adviser to be the most appropriate source to which to turn for aid—doubtless as liaison to professional sources. Note particularly responses to items 20 and 29.

Students tend to regard as outside the scope of University service their legal problems (note items 18, 24, and 27, and perhaps others in which they feel its role would most likely be punitive (item 23). They do not tend to regard their emotional problems as, *per se*, outside the scope of University responsibility (Items 20, 26, and 29).

Students base their opinions of the Advisory System on a fair amount of information. The mean of 12.2 out of a possible 20 seems high, especially in a sample containing 54 1948 entrants who had been at the University less than a quarter, and who made a mean score of 10.9 themselves.

Responses seem to support a common-sense view of the advisory function. The unanimity of responses on the case-study items seems to indicate very little disagreement among students as to what they want from advisers, and the comments seem to bear this out. They want warmth, understanding and acceptance of their goals and purposes. Where necessary, they want intercession on their behalf. They do not want advisers to play psychoanalyst at them, but it should be borne in mind that an adviser who would do so would not be behaving at all as would a real psychoanalyst attempting to help the same indi-

vidual. The students do not, therefore, reject the concept of therapy, but the possibility of its being used by someone else to act out his own problems—a good thing for anybody to reject. Most of them accept as desirable, according to responses to item 95, the intercession of the adviser on behalf of an academically successful but troubled student. Many see dangers in this, however, and a few react strongly against it.

There is a remarkably conservative, and, if one may say so, uncritical and middle-class orientation of student values, related to religious and personal freedom. There is strong, and, again, perhaps uncritical identification with the College, its purposes, and what they conceive to be its mores. (2) Evidently, students do view the system as a part of the total educational service of the institution, and expect its functions to be modified in the light of, or perhaps even determined by, the institution's purposes.

Internal criticism of the inferences made is possible, though laborious, by a statistical analysis of differences among the sub-groups on relevant items. For example, if one reason the data cited fall into the pattern observed is that students view the Advisory System realistically, and do not attribute to it psychoanalytic functions, one would certainly expect of younger and less experienced students that they would make choices indicating somewhat more dependence than the rest of the group. An examination of item 29, response B, reveals that 56 per cent of students in the 11th and 12th grades, aged 18 and under, regard the adviser as the most appropriate University official to approach with this highly clinical problem, as compared to only 37 per cent of the remaining group. This difference is significant at the 5 per cent level, yielding a critical ratio of 2.1. On the other hand, 22 per cent of the younger group choose response D, as compared with 31 per cent of the remaining group, which is not a significant difference, this, too, is perhaps explicable, since a significant difference on this response would indicate positive disillusionment with the system with growing independence and maturity, which presumably does not occur. On item 95, 10 per cent of the younger group choose response A, as compared with 24 per cent of the remaining group, a difference yielding a critical ratio of 2.3, and to

be expected in view of the greater independence of the older adolescent or young adult. Fifty-six per cent of the younger group, as compared with 48 per cent of the remaining group, however, choose response D, a difference in the expected direction, but not significant and not sufficient to constitute evidence that the older group repudiates the assistance of the system in solving personal problems.

If the results obtained by applying the instrument described at the University of Chicago are representative, then, it seems that while students feel that they need warmth and understanding and that the University is obligated to provide help with personal problems, they are not likely to misuse or overburden the source of such help. They will, in general, take as much as can be given of what they need. The more psychological insight which the Advisers in a system possess, and the more clearly the system defines its scope to include service with personal problems, the more students will expect of it and use it. Some, however, will become frightened and hostile, and most expect enough initiative to be left to them to permit them to feel respected, rather than manipulated.

REFERENCES

1. Blom, Peter. *The Adolescent Personality*. New York: D. Appleton Co., 1941.
2. Kelly, Janet A. *College Life and the Mores*. New York: Bureau of Publications, Teachers College, Columbia Univ., 1949.
3. Krugman, Morris. "Orthopsychiatry in Education." *Orthopsychiatry*, 1923-1948, Lawson G. Lowrey (Ed.) American Orthopsychiatric Association, 1948.
4. Munroe, Ruth L. *Teaching the Individual*. New York: Columbia University Press, 1942.
5. Rauchenbush, Esther. *Psychology for Individual Education*. New York: Columbia University Press, 1942.
6. Rogers, Carl R. *Counseling and Psychotherapy*. New York: Houghton Mifflin and Co., 1942.
7. The College of the University of Chicago. *If You Want an Education*, n.d. (Public statement, released 1949).
8. Zachry, Carolyn B. *Emotion and Conduct in Adolescence*. New York: D. Appleton Century Co., 1940.

THE ROLE OF STUDENT GOVERNMENT IN THE STUDENT PERSONNEL PROGRAM

BROTHER LOUIS

Dean, St. Mary's College, Winona, Minnesota

THERE are so many different interpretations attached to the term, student government, that it would seem almost necessary to open this discussion with a definition of it. However, I am going to sidestep that responsibility, and hope that my definition of student government will gradually be recognized from what I have to say about it. I doubt that a concise definition could be given which would not be subject to various interpretations. And, so, for the present, by way of preliminary explanation but not as a complete definition, I will say only that when using the term, student government, I have in mind a student organization composed of the highest elected officers of the student body, having very definite and real responsibilities for all student life and student activities on the college campus, and working in close conjunction with faculty, student body, and administration. My comments will be directed towards three main points: (1) the place that student government should have in the total personnel program of the college, (2) the functions it should fulfill, and (3) the conditions that are necessary in order that it can effectively carry out these functions.

Student government must be an essential and integral part of the total personnel program of the college because it is the one means for accomplishing those aims of the personnel program which are related to and achieved by group living and group activities. While other areas of college personnel services are concerned primarily with the student as an individual, the area of student government is concerned with the student as a social being, in relation to both the college community and the other social environments in which he will live. It is the means for unifying all efforts of the college toward the education of the student as a social being. Since, then, it is

one of the personnel services, it should have the same recognized status, the same prestige, and the same freedom to operate within its sphere of responsibility as, for example, the health service. It should also, by its very nature, have the same all-pervasiveness with respect to the whole college program as the counseling services.

This implies that the program of student government comes directly within the scope of responsibility of that administrative officer of the college who has general charge of all student personnel services. On most campuses this would be the Dean of Students. It also implies that the authority and responsibility which the student government has are delegated and not absolute, that is, delegated by the administration to the student body to be exercised by the elected officers of that body in accordance with a constitution accepted by the student body, the administration, and the faculty. If correctly understood, this places no real restriction on the student government, since, just in the same sense, the authority of the Dean of Students or the Dean of the College is delegated and not absolute. The crux of the matter is really the good judgment of the higher administrative officers of the college, who can either make or break the program of student government according to their attitude toward it. If restrictions are imposed to such an extent that there is no possibility that the student officers will make mistakes, then the program is doomed to failure.

Briefly, then, the student government is an integral part of the student personnel program of the college, and it has a delegated authority which comes to it through that administrative officer who has been charged with the general responsibility over all personnel services.

In order to merit and maintain the status that it should have, the student government has several important functions to fulfill. The major ones, I would classify as follows:

1. It should have the responsibility for the operation and control of all student organizations of the college campus.
2. It should have the responsibility for promoting, organizing, and directing what might be termed "all-college" functions and programs, that is, those which involve the whole student body and not just one particular organization or group.

- 3 It should have a definite responsibility for the formation of policies concerning all student life and student activities of the campus
- 4 It should provide the means for achieving mutual understanding and close cooperation between students, faculty, and administration

Each of these functions needs some explanation. Under the first of them, the responsibility for student organizations, would come the reviewing and approving of constitutions, the setting up of standards, the auditing of books, the supervision of social and other affairs of these organizations—such as dinners or dances—the education of officers of these organizations, the authorization of student concessions, the supervision and control of student publications and student bulletin boards, the fostering of wide student interest and participation in the various campus organizations, and so on. Much can be done by the student government toward the education of officers and members of these organizations in their duties and responsibilities. Sponsoring and directing leadership workshops open to all students is one, providing consultation services is another, and developing brochures giving helpful suggestions is a third. Two such brochures which I recently received from Washington State College are excellent examples of what can be done. One is called "My Chairman" and explains the rules of order concisely, yet adequately. The other is called "Officers' Blueprint" and has many good suggestions and recommendations.

The second general function of student government stated previously is the responsibility for what we called "all-college" affairs and programs. This would include, first of all, "all college" social functions and affairs, such as dances or similar functions, which are common to all colleges. Other types of programs would perhaps vary from campus to campus. As illustrations of those for which the student government can assume full or partial responsibility I would suggest the following campus activities for the annual homecoming, field days, the relations of inter-college student associations with the student body, parents' weekend, the orientation of new students, and convocation programs. We can include here, also, the sponsoring of student forums and inter-college conferences, on student problems or on pertinent topics of the day. If the college

has a student union with a union board to direct the activities centered there, this also, I believe, should be placed under the general responsibility of the student government.

The third general function of student government concerns the formation of policies. Conditions of student life and the operation of student activities are certainly of great concern to the student body as well as to the faculty and administration, and policies concerning them are much more effective if the students have a voice in their formation. The formation of such policies should be a cooperative or joint responsibility of students and faculty. (The term "faculty" will sometimes be used loosely here to include both faculty and administration.) Hence, a joint student faculty committee, meeting weekly, is a practical necessity for this purpose. Such a committee has been set up on a number of campuses. The committee should be appointed by the president of the college, with the student members designated by the student government. Its purpose is to draw up policies governing student life and student activities at the college. It should have the same recognized status as do all of the other committees appointed by the president. The student government should have the responsibility not only of designating the student members of this committee, but its approval, as well as that of the faculty, would be required before any of the policies proposed are accepted. It can also assist the committee by the recommendation of points for incorporation in the policies to be proposed.

Considering, now, the last of the stated functions of student government, it is obvious that a college educational program can operate effectively only in an atmosphere of mutual respect and understanding and cooperation between the three major groups which compose the college community, students, faculty, and administration. There must be an opportunity for free discussion and interchange of ideas between all three. The student government furnishes an effective instrument for achieving this desired result, if channels are provided for direct approach to each of these major groups.

For contact with the student body, a necessary means would be a student convocation, monthly or oftener, conducted entirely by the student government. Contact should also be main-

tained through the medium of student publications. Other means would be the holding of meetings open to the student body, and having an office and definite office hours when students can come to present and discuss problems, questions, and suggestions.

For contact with the faculty and administration, there is, first of all, the faculty or administrative adviser and the student-faculty policy committee mentioned earlier. Other means will vary according to local conditions. On our own campus several changes have been made as our program progressed. Originally we had regular business meetings of a college council, composed of the student government and a committee from the faculty and administration, of those directly concerned with student problems. While this body had other functions, our main idea was to educate the student and faculty groups to an understanding of and respect for the viewpoints of the other. However, so much has been done to modify viewpoints and attitudes of both students and faculty as to make the lengthy discussions we used to have, unnecessary. Questions or problems which arise now, are taken up directly with the proper administrative officer concerned or with the student-faculty policy committee, and settled promptly and satisfactorily. Since business meetings of this college council are no longer necessary, we are changing this year to informal luncheon meetings in order that the two groups will still keep in close touch with each other and so safeguard the harmony in attitudes and thinking.

Originally, too, we held separate meetings also of the faculty committee which was part of the college council. This would be poor practice if the intention had been to form a united front to present before the student officers. Our intention, really, was to modify the somewhat extreme viewpoints of one or two of our members, and this enabled the joint meeting with the student group to proceed more smoothly as a result. We have since discontinued this separate meeting of the faculty committee.

Another means which makes for good relations is for the faculty and administration to consult with the student government even on those problems and policies over which they

retain final decision. If changes in policy are explained beforehand, together with the reasons which influenced the decision, much better understanding and cooperation on the part of the students can be achieved.

A program of responsible student government often requires much patience in the beginning. At first there is quite likely to be a distrust and sparring for advantage which delays progress. There is also, at first, a tendency on the part of students to be preoccupied with very petty problems and to ignore the really important ones. This is not their fault since they have not had any previous education in this respect. However, with tact and patience this difficulty can be overcome, and the final result is worth the effort.

It may seem that I have completely ignored the area of student conduct and student discipline in relation to student government. This is not my intention since I regard it as being included under each of the four major functions discussed. Through the student-faculty policy committee, for example, the student government has a very definite voice in formulating policies regulating student life and student conduct. And the responsibility for activities of student organizations or for the whole student body entails a responsibility for student conduct in connection with such activities. Further, it is my conviction that the student government can also assume the responsibility for supervision of student conduct in residence halls. I have not placed this as one of the major functions of student government because it seems to me that too great a stress on the area of student discipline implies a negative rather than a positive approach, and can lead to a neglect of other important areas. It also surprised me at first that for the most part, the men students, at least, do not care to assume responsibility for the supervision of student conduct in residence halls unless there is considerable dissatisfaction with the way this supervision is being handled. At a conference on student government which our students sponsored earlier this month, and which was attended by delegates from about twenty-five colleges, this question was discussed at some length. According to the report I received, only one of the men delegates was strongly in favor of having the student government assume responsibility for the supervision of student conduct in residence halls. All were very

anxious to have a voice in determining the policies, but, in general, did not want to go beyond this

As a final point, I would like to comment on some conditions which are necessary for the effective functioning of any program of student government. I will pass over the necessity of having the authority and responsibility of that body clearly defined, as being too obvious to need comment. Outside of that, the most essential condition is that there be good relations between the student body and the faculty and administration. If the faculty lacks confidence in the student group and its representatives, if it is unwilling to take time to discuss fully with them questions and problems of mutual concern, thus ignoring the educational possibilities this affords, then the program is foredoomed to failure. The result will be that in the students' minds the college community will be composed of two rival factions, students versus faculty, each struggling against the other. The administration must take all possible means to prevent such a situation. Some means which can be used have been indicated earlier, but even these will fail unless the viewpoint of the faculty and administration is one of respect for and confidence in the student group.

A second necessary condition for an effective student government program is provision for insuring continuity of policy and the education of student officers. If each newly elected group must start from scratch, there will be no appreciable growth. One practice which is good and is also quite common, is to have the new members elected early enough so that they can sit in at all of the meetings of the present members until the time comes for them to take office. Another means which our own student government uses seems to me to be even more fruitful. In late Spring they take the officers-elect away to a camp for a two-day orientation. Several members of the administration and faculty are invited also. The time is spent in discussing with the newly elected officers, the problems which came before the student government during the past year, the solutions arrived at, the policies established, and the projects and plans for the future. The specific purpose is the education of the new officers in their duties and responsibilities. It really works.

A comment might be made also on the size of the student

government body. This also is a contributing factor to effectiveness. If the group is too small it becomes ineffective, or is in danger of becoming the tool of pressure groups. If it is too large it becomes unwieldy, and tends to lose the "esprit de corps" which should characterize it.

This brief discussion of student government does not pretend to exhaust the subject. I have attempted simply to explain those points which have impressed me most strongly when working with students. Many others could undoubtedly be added. The one general impression I have from my own experience with students and their officers is their willingness and their ability to show a strong sense of responsibility, to be mature in their judgments and to understand and discuss intelligently the problems involved in dealing with people. Capitalizing on these student traits goes a long way toward achieving our educational objectives and toward developing the educated leaders we want our college graduates to be.

STUDENT PERSONNEL WORK AND THE NATIONAL STUDENT ASSOCIATION

GORDON KLOPF

Chairman, National Advisory Council, N S A, University of Wisconsin

I BRING you greetings from the National Staff and the National Advisory Council of the United States National Student Association. It is gratifying to both the Council and the Staff that the American College Personnel Association has always had a place for discussion of the National Student Association on its convention program. This is perhaps rightly so—for who should be more concerned about a program affecting students in over three hundred colleges than the personnel workers in those colleges? Before we explore the role of the college personnel program and staff persons in relationship to the National Student Association, let us observe what NSA is doing and what its future plans are.

In studying the objectives and programs of NSA, we find a great emphasis given to the importance of training students for citizenship. To serve this end, NSA is urging the development of the campus as a community—a community of students, faculty, administrative, clerical and service staff, as well as regents, trustees and alumni. To make a community philosophy function, students must be represented on major committees and boards—particularly those which affect student life. If the campus or educational community is to be an educational experience and a training for citizenship, it must be more real than something we put on like “Sunday go-to-meeting clothes.” In all phases of campus life, an opportunity must be provided for democratic processes to function.

Few institutions have given students an opportunity to play a part in the academic planning program. If we are to give the student a realistic experience in democratic community planning, it is essential that we break down some of the distinction between the student and the teacher. As Harold Taylor, Presi-

dent of Sarah Lawrence College, says, "Education is not something done to students, it is something students and teachers do together." Educational planning has been chiefly the imposing of the academician's point of view upon the student; NSA urges greater consideration of the student's point of view. President Blanding of Vassar says, "Student opinion concerning matters which are considered to be the chief responsibility of the administration and faculty is extremely important, particularly if presented in a thoughtful, constructive and responsible manner." In California we find President White of Mills College "deploring the lack of faculty respect for student opinion." He thinks NSA should accept the challenge to do something to generate a greater respect on the part of faculty and administrative personnel for student opinion.

NSA has developed, as many of you know, a program of student-faculty evaluation. The first edition of the program describing student-faculty evaluation sold out shortly after it was issued. Copies of the second edition are still being ordered in large quantities. This publication contains basic principles, forms, and procedures which can easily be adapted to the local institution. Students are deeply concerned about improving instruction—and who knows more about the instruction they are receiving than they do?

Among the pioneer programs in student-faculty evaluation were those at the University of Michigan, University of California and the University of Wisconsin. Recently, one of the departments at Wisconsin had an assembly with both faculty and students in attendance to evaluate the role each played in the instructional work of the department. These and similar programs have been motivated by the National Student Associations's work in this area.

An issue which concerns many college administrators is that of academic freedom. At the 1949 Congress, the Association resolved, "That membership in any political, religious, or other organization, or adherence to any philosophical, political, or religious belief does not constitute in itself sufficient grounds for the dismissal of faculty, failure to rehire, or denial of tenure to educators of the United States."

In exploring the role of the student in the government of his

community, the NSA has given impetus to a tremendous interest in student government. A number of excellent publications in the form of booklets and mimeographed program materials have been published by the Association. NSA has stimulated the development of student leadership conferences, student government clinics, and workshops on the local, regional and national level dealing with the role of the student in the governing of higher education.

In promoting the concept of college or university as an educational community, the NSA realizes that most aspects of a community must be governed by trustees, regents, deans, faculty and administrators. It is urging, however, that student opinion and representation be included to a greater degree on committees, councils and boards, giving students the opportunity of expressing their point of view and of having the experience of working with the staff members of the educational community. If we accept the responsibility of higher education as being a training ground for citizenship, we need to think of the institution as a community-structured unit with students as well as staff members as citizens, participating in the planning and governing of the community. It is to this end that NSA is working.

The National Student Association is also interested in developing "concerned citizens" among our students. To implement this objective it has planned an extensive international program. Almost eight hundred students will participate in the tours abroad this year. In providing this travel program, NSA not only saves the American student hundreds of dollars on every tour, but is helping the student to get the maximum from his travel experience by making the tour a "study" as well as "sightseeing" program. The student not only sees the Eiffel Tower but learns about the people of France through studying the French language, the history of the French people, and their customs, while on board the ship taking him to Europe. In France he meets with student groups and with people of France other than the "Cook's Tour Guide" type of individual. When the student returns to his campus, NSA has urged him to try to give other students some means of benefiting from his experiences abroad. NSA has also developed a program of

work camps and has done a great deal to bring displaced persons to the American campus. It is constantly working with other agencies in the international field. Shortly, you will see on your desk a copy of *Youth and UNESCO*, a new publication which has been published by NSA and UNESCO. NSA has been a vital part of the program of the World Student Service Fund. Many of you may have heard about the expanded role of World Student Service Fund in the program of international education. NSA has been consulted on this program, and, as it takes shape, NSA leaders will be involved in its implementation.

The National Student Association has also urged colleges to permit political activities on campuses, including the permission for speakers of all political views to appear, and the development of political organizations on the campus. I think it might be said that the NSA agrees with Robert Hutchins when he says,

The policy of repression of ideas cannot work and never has worked, the alternate to it is the long difficult road of education, to this the American people have been committed. It requires patience and tolerance, faith in principles and practices of democracy, faith that when the citizen understands all forms of government that he will prefer democracy and that he will be a better citizen if he is convinced than he would be if he were coerced.

The program of the Association is interested in developing a "socially concerned" student. All phases of the National Student Association's Program are aimed at providing experiences in inter-group and inter-personal understanding. The 1949 Congress certainly served to illustrate the importance of students who represented different backgrounds and points of view working together when a sub-commission dealing with a debatable statement of policy refused to present a final draft until the students holding an opposing point of view were consulted and placed on the committee. Through the regional, state, and national conferences, students of all races, religions, political backgrounds, geographical regions, social and economic status have an opportunity to work together. The national congresses have taken definite stands on discrimination in student groups

and have asked member student governments to prohibit organizations which discriminate against groups of individuals. To help implement the best in human relations in the Educational Community, the Association has recently published a booklet, *Human Relations in the Educational Community*. This, as well as many other program materials issued by the Association, will give students, faculty members, and administrators suggestions for meeting the challenge so ably stated by President Charles S. Johnson of Fisk University that

Unless the American people solve the racial issue they face a national defeat from within through loss of faith in their very reason for living. We cannot rest now, or turn back the tides, or settle the crucial issues by comfortable compromises. We can either be courageously righteous in our belief in ourselves, or adopt an ideology and way of life to fit our inseparable sins.

The National Student Association is also concerned with the economic welfare of students. As many of you know, it has developed a Purchase Card Plan which has been successful in many educational communities. The staff realizes that the plan is not workable in every community and has developed other means of helping students to meet their economic needs. The NSA is distributing program materials concerning cooperative stores, housing and eating groups. At the 1949 Congress, it approved by an overwhelming majority the need for federal scholarships to be awarded on the basis of need and ability and, recently, the national staff participated in a conference sponsored by the American Council on Education in the drawing up of a bill for federal scholarships to be presented to Congress.

In concluding this section, which has given you a brief picture of the program of the Association, I wish to say that I agree with President Harold Taylor of Sarah Lawrence College that "a lethargy is present in the American student body which has resulted from the fact that our college and university administrators and faculty have not given sufficient encouragement and opportunity to the participation of the student in the total life of the campus." The National Student Association is three years old, and I am sure you will agree with me that it has done much to encourage student participation in

the total life of the campus. Part of the success of the Association, however, is in your hands

It is important this morning that we also examine the structure and administrative procedures of the Association. The most frequent question asked of the Advisory Council, now that administrators are assured the Association is not overrun with "fellow travelers," concerns the matter of cost. Many of us will admit the cost has been high and have urged the Association to study the possibilities of reducing membership fees. Since the membership has increased, dues have been reduced, and are going to be reduced to an even greater degree. However, it is important that we compare the cost of membership in the National Student Association and the cost of student government to other activities on campus. There is hardly a college that does not spend more on debate and forensics, with relatively few students participating, than they do on student government or membership in the National Student Association. The experience of a student who attends a Conference or the Annual Congress is just as important to that student as his participating in a debate tournament or a regional forensic contest. Are American institutions as willing to spend money to educate for citizenship as they are willing to spend money to buy band uniforms, train baton twirlers, debaters and athletes?

I believe that, basically, the problem with the National Student Association is not the three cents it costs each student on the campus to belong, but rather lies within its structure. The organization nationally consists of the Student Congress which meets annually, the National Executive Committee, composed of Regional Representatives, which meets between Congresses, the Staff Committee, and Regional Organizations. The weakness in its structure lies in the Regional organization and in the local campus channeling. On your local campuses, the person who should be most concerned with the program of the Association is your student government president. Because of the complexity of his job, he may have assigned the channeling and coordination of the NSA materials to a special committee, commission or coordinator. However, my experience with local campus structures indicates that the closer the president is to

the NSA program, the more he reads and channels program materials to proper committees, the better the purposes of the National Association are being served. Let us take, for example, the recent material that Ted Perry, the Vice President in charge of Student Life, has sent to the Student Government President concerning campus social and recreational programming. Ted has developed an excellent collection of materials concerning both formal and informal campus recreational programs. When the student government president receives this, he should immediately forward it to the Campus Social Committee, the Union Dance Committee, the Dormitory Social Committee, or whatever Committee or Board is concerned with planning campus social activities. He might also refer it to the Dean of Women or Men, the Student Activities Director or the Dormitory Social Director. I cannot urge you as personnel people too strongly to be sure the material that the National Student Association distributes is read by the people who should be concerned with the particular project. The campus that permits these excellent suggestions to lie on the student government president's desk is certainly not getting its money's worth from membership in NSA. Again, I say it is not the cost factor of the Association itself—it is the inefficiency of our own student leaders in channeling the NSA program material.

Another factor is that of leadership in the Association. It has frequently been said that "students will be students" and cannot accept the responsibility of administering a national organization of the scope of the National Student Association. I think it is important that we become convinced, along with Dean De Vane of Yale University, that it will not do to underestimate the abilities of our young people, and that, if the organization has not enough in itself to assure its continuation, it ought to die. I think we also agree with the college president who said that, "We do not want to see an aging secretariat grow up in NSA." However, I think we have to realize that the mature leadership of the post-war veteran student body is no longer present. We all realize that the American student body is not as mature as the student bodies of two years past. Our job as personnel workers is to be sure that we encourage our local NSA programs and write to the national officers to give

advice and suggestions. Administratively, the Association is meeting the problem of continuity of leadership through having several of its officers run from February to February and others from September to September. If ever there was a need for a National Student Association, to develop concerned citizens, it is at the present time.

Last of all, I would like to refer specifically to the role of the National Student Association and the personnel worker. The Association is established to achieve many of the same objectives in which we, as counselors of students, are interested. If all its printed materials, its hundreds of answers to individual letters on local problems, and its regional and national conferences and congresses are fully utilized by your campus, the Association can help you achieve the objectives of your personnel program. The local, regional and national leadership, however, needs your help. It's up to you to carry out the lines of the song, "Accentuate the Positive and Eliminate the Negative." I have attempted, today, to mention just a few of the positive contributions and significant objectives of the National Student Association. Again, I say it is our job, as personnel workers, to be informed about them and to help the student to accentuate them.

CONTRIBUTIONS OF THE STUDENT UNION TO THE TOTAL PERSONNEL PROGRAM

(An Abstract)

DONOVAN D LANCASTER

Director, Moulton Union, Bowdoin College, Brunswick, Maine

AMONG the many personnel services that characterize the contemporary American college, the student union is a comparative newcomer. Student union goals may be expressed simply:

- 1 To help provide a recreational program for the student body
- 2 To reduce the cost of going to college by supplying inexpensive recreation
- 3 To further fellowship and understanding by providing an opportunity for students of different races and social backgrounds to meet on an equal footing.
- 4 To promote the personal development of the students by bringing to the union the best in the arts and by giving the student an opportunity to participate in gracious social gatherings
- 5 To provide a situation where students participate in self government and learn to cooperate with others and to take responsibility
- 6 To unify the campus, large or small.

The bronze plaque at the front entrance of our own Bowdoin Union says, "Here the fires of friendship are to be kindled and kept burning."

What about the administration of these organizations called student unions? The most successful unions in this country are located on coeducational campuses and are housed in coeducational buildings. We have seen the utter folly, even within the last decade, of building separate plants for men and women at opposite ends of the same campus. It sounds amusing to us today, but it is also tragic.

Now the driving force within any student union is the director. While about 85 per cent of our union directors are men, some of the directors of large successful coeducational unions are women. I know a number of coeducational unions with excellent women directors. I shall not go into the merits of this situation, but I shall take advantage of my position here and refer to the director as he. The union director largely determines the union goals because he is on the job day after day and because he is the manager of the building. He should be in charge of all personnel, directly or indirectly, within the union. Otherwise, many times his hands are tied. It is difficult for some college presidents and business managers to see this point.

The union director should be advised and assisted in policy making by a faculty-student board. Faculty and staff members indispensable on such boards include deans of students, student counselors, directors of student activities, teachers of psychology and allied personnel workers. Here is the great chance for personnel officers to make their influence felt. On the other hand, in a smaller institution like my own, the union director also serves on various boards for the deans' offices. This is a desirable interlocking arrangement.

Student members are also indispensable on policy-making union boards. Here, as almost nowhere else, the undergraduate tries his wings in student government and organization. He has a building, a workshop, a program to direct. Here is democracy really at work.

The program of the National Association of College Unions for nearly ten years has contained papers describing the job of the union director and his responsibilities for coordinating his program with that of other personnel offices. If he is not doing so, it may be because he has been charged by the university with the task of making a multi-million dollar building pay for itself and that he has little time for anything else. I am sorry to say that I think there will be more rather than less tendency in the future for university officials to put pressure on the financial rather than the personnel considerations in the direction of college unions.

During the next few years enrollments will decrease and student personnel staffs, including student union staffs, will likely

be reduced. The directors of the many new union buildings that have been built recently, or those in the process of completion, are likely to face vexing financial problems. Every effort must be made, therefore, to avoid overlapping in services and to coordinate the union programs with the larger university or college personnel programs. Listed below are some important steps that might be taken:

1. Centralized recording of the social and recreational interests of students' might overcome the expense of duplicate records
2. The student union organization might contribute more effectively to the freshman orientation program of the institution
3. The creative arts program of the union might become an important laboratory for academic instruction in these areas as well as a setting where students may acquire recreational skills or vocational tryout experiences
4. The student union organization must go far beyond the building itself. Returning veterans who have experienced the possibilities of successful student union organizations on various campuses have often established on their own campuses effective programs without buildings or with very inadequate facilities

For those of you who wish to pursue the whole subject of the student union more fully, I suggest that you consult *College Union—A Handbook on Campus Community Centers*, by Edith O. Humphreys. This is the most exhaustive study of student unions made in America. If you are planning a new union building the National Association of College Unions stands ready to help you. Inquiries should be addressed to Edgar A. Whiting, National Secretary at Cornell University.

MAJOR ISSUES AND TRENDS IN THE GRADUATE TRAINING OF COLLEGE PERSONNEL WORKERS

W. W. BLASSER

Specialist for Student Personnel Programs, U. S. Office of Education
and

CLIFFORD P. FROHLICH

Specialist for Training Guidance Personnel, U. S. Office of Education

WHAT are the major issues and trends in the graduate training of college personnel workers? We will not attempt in this brief paper to present a definitive answer to this question. Our purpose is to try to stimulate discussion through a rather arbitrary selection of issues and trends.

A first step in considering this problem of training is to answer the question, "Who are personnel workers?" We must have clearly in mind who we are training before we can talk about what kind of training they should have. During the past fifteen years we have had quite a flowering of books, articles, speeches and committee reports identifying student personnel functions, services and workers. Despite considerable variation and, at times, conflicts in our literature and in our practice, we now seem to have a fairly common understanding of the general scope and functions of personnel workers. We generally agree that instruction, business management, public relations and maintenance are not personnel work. But when we get to specifics, we find the first issue to raise. The following Committee publications will illustrate the point at hand.

In 1937 the American Council on Education brochure, entitled *The Student Personnel Point of View*, included health as one of 23 student personnel functions. In the 1949 revision of this brochure, health functions were again included among the 17 basic elements of a student personnel program.

Also, in the 1948 report of the ACPA Committee on Professional Standards and Training, it was recognized that health services were one of the student personnel functions. Yet the

Committee sidestepped the issue of the training involved by commenting as follows

Two types of personnel services are included in the list of functions with which we started, but not in the special training recommendations. The first of these consists of positions for which recognized standards are already set by some accrediting agency. Physicians and nurses in the health service would fall into this category. . . it would seem to be advisable to let the persons in the administrative position under whom these activities fall set up standards for them which are in accordance with the goals of the program of the particular institution.

The issue, then, is whether or not the occupational group of personnel workers include all those who perform personnel functions. Are nurses serving college students personnel workers? If they are, should they be trained as personnel workers? Or is their primary allegiance to the occupation of nursing? We have, on the college scene, many persons whose primary techniques are derived from other professions, such as nurses, social workers, speech correctionists, physicians and clinical psychologists. In any list of personnel functions, the services rendered by these individuals are usually included. Is personnel work really as inclusive as the 23 functions listed by the ACE would lead us to believe? Or, is personnel work a small nucleus rendering a unique service which is concerned primarily with the individualization of education by means other than instruction, maintenance and administration?

This issue leads us clearly to the next, concerning the professional status of personnel workers. Is our occupational group really a profession? Darley and Wienn carefully considered this problem and proposed eight criteria, against which the occupational group could measure its degree of occupational professionalization. They concluded that, as a whole, student personnel work falls short of professional status, by all of their criteria save one, namely, we do have a body of specialized knowledge and skills. Of course, the question of whether or not we are a profession is largely academic. For the purpose of this discussion, it is important to recognize that as an occupational group, although somewhat ill-defined, we do seem to have a set of unique skills and a body of knowledge. This,

we believe, is one of the most important reasons that we, as personnel workers, have for considering today the training of personnel workers. If we did not have something unique, then we could leave our training problems up to other disciplines. When we needed workers we would then recruit from other disciplines.

A thoughtful discussion of the uniqueness of our training is found in the ACPA report just referred to. This report stressed that

Since all personnel workers have as their central aims the welfare of the individual student, and his adjustment to the college situation, both in and out of the classrooms, it has seemed to us that training for all should be built around a common core. This should involve information with regard to individuals as individuals, and as members of groups. It should also include the development of skill in identifying individual needs and problems, and handling interviews and group leadership situations constructively.

The common core was then outlined in terms of course work, along with a general recommendation about the need to include supervised experiences. In addition, the report spelled out five rather specific groupings of personnel occupations, and indicated the desirable training recommendations for each group.

So much for this forward-looking report. We need now, for the purposes of this paper, to make a rough and arbitrary classification of the majority of training programs available today.

We shall admit readily that a particular program may not fit perfectly into one of these categories. But the classification does serve to highlight certain issues. First, there is the "if-some-is-good, more-ought-to-be-better" type of program. This training has its primary orientation in counseling, an applied branch of the science of psychology. In this program, *levels* of personnel workers are recognized. Those with bachelor's degrees in psychology are considered capable of working as placement interviewers in the College Employment Office. They can be resident dormitory counselors while they work on their M.A.'s, or they may be preliminary interviewers in a Counseling Bureau. At the next level, the M.A.'s can work in the Dean of Students' office, handling simple discipline, loan funds, or they can be counselors in the Veterans Counseling Bureau. By taking more training in psychology, these persons may earn

a Ph D. They are then eligible to move up to the top level. Here they can become the Dean of Students, or if they are so inclined, can get into the teaching end and train more personnel workers. This program is characterized by adding more and more training in counseling, upon the apparent assumption that the higher the counseling skill the better the personnel work.

Now there is a second type of training program, namely, "to-each-his-own-specialty" type. These programs recognize *areas of specialization*. By enrolling in this type of training program, students can be trained as vocational counselors. Their training may duplicate, in part, that of a personal counselor, but the training program will tend to accentuate differences, special skills, rather than skills common to all personnel workers. Earlier in this discussion we pointed out a pertinent example of this "area of specialization" approach in which certain special groups for which standards were set by some other profession, were accepted as personnel workers. It is our belief that this type of program is based upon the assumption that there is not a body of specialized knowledge and skills in personnel work. Rather, the personnel functions are a conglomerate of occupations, under a single banner, and not a single occupation with a variety of specialties. If we carry this belief to its logical conclusion, each of us as personnel workers would have our primary home in some other professional land. In fact, we would think of ourselves as psychologists, or dietitians, or vocational counselors, who just happen to be working in a college.

In some quarters there is strenuous opposition to this second point of view. And this opposition has been the motive power behind the establishment of a third type of training program, namely, "be-a-generalist, be-an-educator" type. The training program is quite logically designed to provide a broad basis in education.

The students get this in courses in the principles, history and philosophy of higher education, and in methods of educational supervision and administration. Primarily, this point of view stresses the *setting* in which college personnel workers find themselves.

This point of view, while it recognizes the importance of the

setting, fails to recognize the unique body of knowledge and skills which personnel workers should possess.

These somewhat facetious and critical descriptions of training programs highlight three of the important elements which we believe should be characteristic of every training program. College personnel training programs *should be* designed to provide for *levels* of specialization, *areas* of specialization, and the *setting* in which the students will work. We recognize that real problems will arise in organizing a program which takes cognizance of all three elements. These problems can be pointed up by considering two types of workers which are now ordinarily found in college personnel programs, namely, the counselor and the college nurse. At the present, the college nurse is trained under medical auspices. Her status is accepted by the medical profession, and to it she owes her primary allegiance. She is truly trained to serve in her area of specialization. However, she receives little, if any, training in the specialized knowledge and skills of personnel work. Likewise, her training for service in the educational setting is neglected. Under the program proposed, this nurse would receive her training to the full competency which she now has in her speciality, but, in addition, she should receive the common core of training which was specified in the 1948 ACPA Committee Report, previously referred to. She should be trained so that she understands the goals and objectives of educational institutions and of the personnel program in them, and of her role in that program. With such training, this nurse would not see students as a parade of physical disorders, of stomach cramps, and headaches, but rather she would keep in mind other possible aspects of the student's adjustment. The student whose stomach is upset because of fear of failure would not only get bicarb, but he would be referred to the Counseling Bureau.

What about the counselor? If he were trained in a program which fully recognized the necessity of these three elements, he, too, would be a more efficient personnel worker. Instead of striving for status as a sort of junior psychiatrist, he would clearly recognize the uniqueness of counseling as a service in the educational setting. He would recognize its contribution and its place in the development of the total educational pro-

gram. He would recognize the levels of counseling skill that are possessed by his personnel work colleagues and by his fellow educators. And, by the very recognition of levels of counseling skills, he would build respect for himself among other staff members on the campus. If, for example, he really believes that college faculty members have a role in the counseling process, then he would make intelligent use of such counseling skills as they might possess. If he really believed that all of the staff members of the institution had a common goal and purpose, that of enabling the individual student to achieve maximum learning from his total college experience, then he could join with them under the banner of personnel work.

These two illustrations have been cited to point up some of the difficulties that are involved in organizing a training program for college personnel workers. Recognizing levels of specialization and areas of specialization, and the nature of the educational setting, will do much to produce an adequate training program. The issues, then, can be stated simply as: Are all persons who perform personnel functions to receive at least a minimum of training in personnel work? If they are, how shall that training be organized so that each may attain competence in his specialty? And how can that training be organized to provide for workers at various levels of competency? Finally, how can the training be planned so that students become familiar with the setting in which they shall work?

It is easy to raise issues when you are not charged with the responsibility of providing the answers. We find it more difficult to fulfill the second part of our assignment today, that of identifying trends in the training of college personnel workers. The spotting of trends is frequently a combination of limited observation and pious, hopeful thinking, a mixture which is not always known to the mixer. Therefore, the following observations are offered without full knowledge of the ingredients involved but with the hope that they may furnish food for the discussion period.

We believe that one trend is an increasing emphasis upon practical supervised experiences, particularly in the training of counselors. These experiences appear to be limited to one or two types within the collegiate institution. A few personnel

workers are urging that trainees be given a wide variety of experiences before being permitted to follow a specialization. Williamson, for example, has urged that counselors be given interviewing experiences in community agencies, business personnel offices, mental institutions, reading clinics, vocational guidance clinics, psychotherapy clinics, and in elementary and secondary schools as well as in the collegiate setting. The internship experience, he believes, should be integrated with the entire period of academic training, and not just tacked on at the end of the formal course-work. A few institutions already are moving in that direction.

Another promising trend appears to be an increasing recognition of the need to analyze training content in terms of actual job function, thus lessening the disparity between one's training and what one actually does on the job. The USES study of educational personnel jobs which was reported by the CGPA Study Commission at this convention should provide a base from which to do more intensive job analysis work. The proposed pilot study of CGPA which was also reported on Tuesday may provide us with techniques and tools by which we can validate training programs against job success criteria.

Recognizing a third trend, some training programs are now providing opportunities for individuals to evaluate and improve their own human relations skills while in training. In short, they are being provided with personal counseling experience, and with group therapy. Please note that we are not advocating that all future admissions officers, counselors and deans be psychoanalyzed while in training. We are simply saying that while they are learning the skills and techniques of personnel work they are also learning to handle their own problems so they do not interfere with the application of those skills and techniques.

A fourth trend appears to be the development of in-service training as a function of student personnel administration, and the recognition of the advantages of coordinating this with the graduate training programs. We have customarily thought of in-service training as a program for graduate students doing part-time counseling in the dormitories, or for members of the teaching faculty who have agreed to work with students be-

yond the boundaries of traditional academic advising. Yet all of us could profit from a well-conceived, long range in-service training program to help us improve existing skills as well as to develop new ones on the job. Here too, the training must recognize and provide for *levels* of specialization, *areas* of specialization and the *settings* in which the jobs are being carried out. The knowledge to be gained from coordination between the full-time in-service and the graduate phases of training should be of increasing assistance in narrowing the gaps between training content and job requirements.

Finally, a fifth trend seems to be increased emphasis upon the *philosophy* of personnel work. The publication of the *Joint Committee on Counselor Preparation*, in which ACPA participated, recommended training in the philosophy which undergirds personnel work. It is clear that a training program needs a sound and carefully defined philosophical base. We believe that there are only a few persons who hold to the mechanistic bag-of-tricks approach to personnel work. Personnel work can never succeed if its practitioners build their strength upon technical knowledge to the exclusion of basic human values.

●

EMPLOYMENT OUTLOOK FOR THE 1950 CROP OF COLLEGE GRADUATES

IWAN CLAGUE

Commission of Labor Statistics, U. S. Department of Labor

COLLEGE personnel workers this year have a particularly challenging task, assisting the largest graduating class in the Nation's history to take their place in the national economy.

About a half million people will receive bachelor's and higher degrees this year, considerably more than last year's record total of 423,500 (The 1948-49 total was nearly one-third higher than the 1947-48 graduation figure and nearly double the pre-war peak reached in 1939-40.)

These large graduating classes, of course, result from the post-war boom in college enrollments, stimulated by the G. I. training program. Enrollments reached a peak of 2,456,000 in the fall of 1949, one million higher than the pre-war record.

The number of students enrolled and the number who get bachelor's degrees will probably drop for several years after 1950, as the veterans move out of college into the labor market. However, the number of master's degrees and doctor's degrees granted should continue to increase for a few more years. And the drop in college enrollments will be only temporary. By the late 1950's, enrollments will begin to rise again, as the first "war babies" reach college age. The long-run trend for a larger and larger proportion of young people to continue their education beyond high school will also tend to push enrollments up.

The great majority of young people leaving college in the near future, like most graduates of previous years, will seek jobs in professional, semiprofessional, and administrative fields. In 1950--probably also in 1951 and 1952--many will be unable to find jobs immediately in the occupations for which they have been trained. There are several reasons for this unhappy prospect. The war-time and post-war shortages in a number of occupations have now been filled. The unprecedented numbers of

new graduates will intensify competition for jobs. Furthermore, there will probably be somewhat fewer job openings for new college graduates in 1950 than in the first post-war years or even last year.

The Nation's economy is currently operating in high gear, and it is likely that employment will continue at about the present high level for the rest of 1950. However, unemployment may increase somewhat, since the American labor force (including both employed and unemployed) is growing at the rate of 600,000 to 700,000 workers a year. This situation presents a challenge to business and industry to utilize fully our increasing supply of potential workers, produce more goods and services, and bring about a rise in our national standard of living. In the long run, I am confident this goal will be achieved; but in the next year, the atmosphere in which college graduates will be seeking jobs is likely to be less favorable than at any time since the war.

Such general observations about conditions in the job market obscure widely varying situations. Prospects are excellent in some occupations, though, in others, graduates will face stiff competition for jobs.

In *teaching*, for example, there is at once an acute shortage of personnel in the elementary schools and a growing oversupply at the high-school level. For the current school year, only one elementary teacher was trained for every three who were needed. On the other hand, 4 times as many students completed training for high-school teaching as were required. This imbalance in supply exists in nearly every State, creating a grave problem both for the schools and for the young people concerned. College counselors can help to remedy the situation by getting the facts on employment-outlook before prospective teachers as early as possible in their college careers.

Other professional fields in which stiff competition for jobs is expected in the next few years include:

Law This profession is already overcrowded and likely to become more so during the next few years. Twice as many lawyers passed the bar examinations in 1949 as in the years just before the war; unprecedented numbers are currently enrolled in law courses.

Engineering In the early 1950's, the number of graduates

will exceed the number of openings in this rapidly growing profession. However, after the next few years, the employment situation for new graduates is likely to improve.

Chemistry. Competition for positions will be keen during the next few years among chemists without graduate training. The outlook is better for those with graduate degrees.

Journalism. The reporting field, always highly competitive, is likely to become more overcrowded in the early 1950's. Jobs will be easier to get with country papers, trade papers, and house organs than with "dailies."

Personnel work. Competition is very keen in this field. Employers are insisting on much higher educational and personal qualifications for positions at all levels than in the previous five or six years.

There will probably also be an oversupply of *business administration* graduates. A surplus of new graduates has already developed in the field of *accounting*.

Liberal arts graduates with specialized training or work experience will find it easier to get jobs than those with only a general undergraduate education.

Fields offering good prospects for new entrants include:

Nursing. A shortage exists despite the fact that there are more nurses than ever before. The demand for nursing service will probably continue to rise.

Medicine and Dentistry. Those able to enter and complete training will have good opportunities. However, competition is very keen for admission to professional schools. Some new schools are opening, more are planned for later in the decade.

Pharmacy. This is a field in which the supply of new graduates has almost caught up with the demand. It is expected that this profession will be overcrowded in the long-run if enrollments in pharmacy colleges continue at present high levels.

Other occupational groups important in health service, such as *veterinarians*, *medical x-ray technicians*, *medical laboratory technicians*, *dental hygienists*, *physical therapists*, *occupational therapists*, and *dietitians* are expected to have good opportunities for a number of years. Women with interest in the medical field will find many openings in most of these occupations.

Social work. Current employment opportunities are excellent in all types of positions. The long-run outlook is good for workers with graduate training, but those with only undergraduate training will face increasing competition.

Psychologists with graduate training, particularly in clinical work, will find good opportunities in the next year or two. However, those with only the master's degree may expect increasing competition. Some psychology majors with the bachelor's degree are having difficulty gaining admission to graduate training.

Many 1950 graduates who have taken training for occupations that are, or soon will be, overcrowded will need your expert help in adjusting to the situation.

For some, the best course may be to take a job in a related field. Thus, many engineering graduates may be able to put their training to use in administrative or technical sales jobs.

For others, the wisest course will be to continue in school for postgraduate work in the same or related fields, in order to improve their chances for employment. This is in line with the long-term trend toward constantly rising standards of educational preparation in many occupations. In engineering, for example, many people with little, if any, college education used to qualify for professional positions on the basis of their practical experience. Now, it is much harder to do this, most openings in the profession are filled by men with bachelor's degrees, and the number of engineers with graduate training, although small, is increasing. The same trend toward graduate training can be noted in many other professions. In addition, the proportion of sales, clerical and administrative occupations for which a college education is required or preferred has been growing rapidly.

Job opportunities in professional and administrative occupations may be somewhat better for graduates who come out of college a few years hence, after the current peak in college graduations has been passed. Employment in the professions has grown rapidly—from $3\frac{1}{2}$ million in 1940 to over 4 million in 1949. It may well increase to more than 5 million by 1960. Employment in administrative occupations has likewise shown an upward trend. In addition, many new graduates will be needed yearly to fill vacancies arising because of death, retirement, marriage, or transfer to other occupations, probably more will be hired as replacements than new jobs. Nevertheless, if college enrollments increase in line with past trends, there will continue to be keen competition for positions in most professional and administrative occupations. This will be even more true if enrollments expand as much as has been recommended by some educators.

Since opportunities will be better in some fields than others, students will need realistic information on employment prospects in different occupations; they should have this before

they enter on a course of training for any field. This information needs to be up-to-date. During the past 8 months the Bureau has been working on a new edition of the *Occupational Outlook Handbook*. The information we have obtained—from industry, organized labor, and professional societies in a great number of fields—underscores the fact that the factors affecting employment trends are constantly changing.

College personnel workers can, of course, do much to see that young people have the needed information available to guide them in making an occupational choice. They can also contribute greatly to a solution of the broader problem of overcrowding of professional and administrative occupations, by helping students to widen their vocational horizons and encouraging them to seek employment in a broader range of occupations.

OUR STAKE IN THE OCCUPIED COUNTRIES

An Abstract

HAROLD E. SNYDER

Director, Commission in Occupied Areas, American Council on Education,
Washington, D. C.

WHY should American educators be particularly concerned with educational developments in Germany and Austria, in Japan and the Ryukyus? What stake, as Americans and as college personnel officers, have you in the reconstruction of the ex-enemy countries and in the rehabilitation of their youth?

The answer is a simple one. It consists of three main points which I believe to be irrefutable.

First, the time has passed if it ever actually existed, when the well-being of American youth can be assured by the opportunities provided in our home and communities, in our schools and colleges. For thousands of our students two terrible wars and the threat of a third even more terrible *have* wiped out and can wipe out *again* all of the benefits of our excellent educational system, all the splendid advantages with which we are trying to provide them. Developments in other parts of the world are of direct and vital significance to all of us, and particularly to our youth.

Second, while the happiness and security of American youth depend upon many factors, it is particularly essential that a concerted effort be made to overcome the effects of the perverted Fascist philosophy and education on the minds of German and Japanese youth. These virile and technically adept peoples must not again be permitted by our disinterest to become sources of infection, infestation and eventually of aggression affecting the whole world.

Poverty and unemployment, frustration and disillusionment, indifference and indecision can once more cause German and Japanese youth to be attracted by the blandishments of totalitarian propaganda, can turn their despair into hatred, can

make them a threat to the security of American youth. Enlightened self-interest demands that we be concerned with aiding the process of educational reorganization and reconstruction and of democratization in the occupied countries.

Third, World leadership has been thrust upon us. By the very fact of our occupation of the ex-enemy countries, we have assumed a very special responsibility for what happens there. In the eyes of the entire world Germany and Japan are proving grounds for the democratic principles which we profess, for the efficiency of our methods, for the sincerity of our motives. We dare not, therefore, fall into the sometimes tempting illusion that these countries can be given identical and equal treatment with all other countries with which we maintain cultural relations. The question is not one of favoring our former enemies. It is obvious that they must not be coddled. But it is equally obvious that if we are to discharge our special responsibilities there, and safeguard our national interests, these countries must continue for some time to come to receive special attention.

PLANS FOR THE NEW INTERNATIONAL CHRISTIAN UNIVERSITY IN JAPAN

An Abstract

MAURICE E. TROYER

Vice President, in Charge of Curriculum and Instruction, Japan International Christian
University Foundation

ALMOST 100 years ago, in 1852 to be exact, the United States officially through her Navy, opened feudal Japan to world trade, industry, and technology. In the 100 years that followed, Japan learned her lesson well, perhaps too well. Today official United States is again in Japan to help with the democratization of her schools and government. Much has already been accomplished in the reorganization of education and government, much remains to be done in educating leaders with a clear understanding of democratic philosophy and processes.

The New International Christian University now being established in Japan, independent of the Allied Occupation Government, has as its dominant aim the education of leaders who will look upon academic knowledge and skills, not as ends in themselves, but as tools useful in working toward: (a) the social order that holds sacred the integrity, worth, and welfare of the individual, and (b) group processes of thinking which provide the basis for enlightened decision and action but which, nevertheless, respect and duly protect the rights of individuals and minorities to pursue their objectives through constructive educational processes.

The University will open in April, 1952, with one undergraduate college and three graduate schools. The major purpose of the undergraduate College of Liberal Arts is to experiment with and demonstrate approaches to general education appropriate to the needs and life of Japan. Traditionally, specialization in Japanese education starts at the high-school level. General education has been unknown in colleges and universities of Japan. It is proposed that the program of general edu-

cation in ICU will include not only natural and physical sciences, social sciences, and humanities, but also agriculture and homemaking, not to prepare specialists in these two latter areas but to bring the contribution of those two areas to the life of the campus. The graduate program includes a Graduate School of Education, a Graduate School of Citizenship and Public Administration, and a Graduate School of Social Work, to prepare leaders for three areas of public service, education, government, and social service.

One of the vice presidents of this new university is to administer and coordinate the student personnel stream of activity—recruitment, selection, admissions, registration, orientation, vocational and educational counseling, clinical services on problems of social and emotional adjustment, health services, housing, student social activities, placement and follow-up.

Educational leaders in Japan have declared that there is no one among the colleges of Japan qualified to handle this position. It will, therefore, be filled by a highly qualified faculty member from one of our leading universities in the United States, who will also head up the program of graduate training in personnel and guidance. This position holds unusual opportunities for pioneer work in the development of new programs, processes, and techniques of guidance in a different, but certainly not new culture and language setting.

Plans for the development of the program for the university are as follows. The beginning faculty, in April 1952, will consist of about sixty staff members. The major function of the new institution is graduate in nature. Since the major function of this University is graduate in nature, faculty members will have completed their doctorate program and at least three-fourths of them will be persons of recognized status in their field. About half of the faculty will be Japanese, the other half from other countries.

A number of the non-American members of the Faculty are to be selected and brought to the United States on fellowships for the academic year 1950-51. About 40 of the faculty members, half Japanese and half foreign (non-Japanese), are to be selected by June, 1951, at which time they will be brought to the United States and assembled on some university campus.

for seven months of planning. This planning session is important. A new faculty representing different institutions, countries, and cultures will need time to think together on the objectives of this university and to build programs and courses which support those purposes. A new system of student records, a library, new equipment—these are all problems for study and development.

In the achievement of these purposes of the planning conference, the faculty will have an unusual opportunity to learn ways of democracy and Christian Brotherhood in their own personal relationships. This is indeed an important pervasive objective of this planning period. The Faculty of this new university should not unduly confuse their students by discrepancies between what they teach and how they behave in relation to each other.

In January of 1952, these faculty members, together with others who will be added in the meantime, will assemble on the campus at Mitaka, fourteen miles out of Tokyo, and prepare to open the university in accordance with the Japanese academic calendar in April, 1952.

In the meantime, classrooms, offices, library, and residence centers for faculty members and students will be provided through a building program costing about three and one-half million dollars. Three hundred thousand dollars have been budgeted for new books and magazines for the library, more than \$500,000 for equipment. Financial plans for the university projected by the Japan International Christian University Foundation in America provide for a reserve of \$5,000,000 to be used as general endowment and a substantial sum to subsidize fellowships and scholarships.

EDUCATIONAL and
PSYCHOLOGICAL



MEASUREMENT

TABLE OF CONTENTS

VOLUME TEN, NUMBER FOUR, WINTER, 1950

<i>Some Statistical Problems in Clinical Research,</i> ROBERT R HOLT	609
<i>The Opinions of Syracuse University Students on Some Widely Discussed Current Issues.</i> N. M. DOWNIE, C. R. PACE AND M. E. TROYER	628
<i>Theoretical Problems in the Selection of Students for Professional Schools</i> EDWARD J. FURST	637
<i>Predicting Academic Achievement with a New Attitude-Interest Questionnaire—I</i> R. C. MYERS AND D. G. SCHWITZ	654
<i>Patterns of Response in Level of Aspiration Tasks.</i> LOUIS D COHEN	664
<i>Evaluation of an Optometric Test</i> A. R. LAUER AND WILLIAM B. MICHAEL	685
<i>The Effect of Client Participation in Test Interpretation</i> PAUL L. DRESSEL AND ROSS W. MATTESON	693
<i>An Experiment in the Rating of Essay-Type Examination Questions by College Students</i> ALBERT ELLIS	707
<i>Relation of Cynicism to Certain Student Characteristics.</i> CHARLES O. NEIDI AND MARION F. FRITZ	712
<i>Change in Teacher-Pupil Attitudes Related to Training and Experience</i> ROBERT CALLIS	718
<i>A Study of Client Responsibility, Counselor Technique or Interview Outcome?</i> CHARLES F. ELTON	728
<i>Recent Publications Received</i>	738
<i>The Contributors</i>	740

SOME STATISTICAL PROBLEMS IN CLINICAL RESEARCH¹

ROBERT R. HOLT²

The Menninger Foundation, Topeka, Kansas

THE title of our round table carries the implication—intentionally, I believe—that clinical research has special aspects which make it somewhat different from most psychological research and which pose certain statistical problems with unusual urgency. Let us first, therefore, take a look at the kinds of things clinical research has traditionally referred to.

The original meaning of the word *clinical*, I am told, was *bed-side*. The clinical practitioner was the practical man who dealt actively and most directly with patients, and thus had to sharpen his sensitivities to everything about his charges that might indicate movement toward sickness or toward health. Research and practice were almost indistinguishable aspects of the same role, for the keen observation that noted similarities in different patients led both to the development of individual therapeutic skill and to the slow amassing of shared knowledge. Thus, to call research “clinical” is to imply first that it has intimately to do with the active dealings of doctors and patients, and that it preserves the clinician’s passion for richness of concrete detail.

As clinically oriented researchers began turning their attention to people who were not obviously ill, it became accepted that research may be called clinical even when it does not deal with patients. The true physician’s approach, which looks toward the whole man in his unique individuality, and the hier-

¹ A somewhat shortened version of this paper was presented at the Symposium “Problems of Statistical Method in Current Clinical Research,” jointly sponsored by the Psychometric Society and the Division of Clinical and Abnormal Psychology at the APA meetings, September 1947. The other participants were P. J. Rulon, R. M. W. Travers, and Daniel Horn; E. L. Kelly was chairman.

² The ideas in this paper were worked out in many discussions with colleagues of the Research Department of the Menninger Foundation, to whom I should like to acknowledge my heavy indebtedness. I am particularly gratified to Roy Schafer and George S. Klein for their assistance and criticism.

archy of scientific values, which puts fidelity to life, keenness of observation and adequacy of concepts to deal with human problems, before the more conventional canons of quantitative precision, objectivity, ready demonstrability, and control of conditions—these came to be connoted by the word *clinical*. It is in this meaning that the term *clinical research* is used here.

Statistical Problems Implied in Three Principal Forms of Clinical Research

Historically the *first* kind of clinical research in psychiatry and psychology followed the lines set down in medicine. Patients whose symptoms were grossly similar would be observed, and the findings collated. From this kind of research in psychiatry, men like Kraepelin reduced the bewildering variety of concrete manifestations of mental illness to a comprehensible schema. The resulting nosology may be creaking and inadequate to the demands made on it today, but it was a real achievement of clinical research in its time.

Research of this kind has never ceased. It is a basic kind of inquiry into the nature of human beings, prerequisite to advancement in other kinds. A psychiatrist who has the good fortune to observe a succession of fetishists, for an example, will publish a summary of the common features of these cases. For research of this stamp, the barest mathematical staples will suffice if any at all are needed: the simple arithmetic of sums, ranges and averages. Even such a refinement as a standard deviation would be out of place if you were working up your observations on six cases of some unusual ailment, such as true paranoia.

A *second* type of clinical research is modelled on the medical experiment of treating patients with a certain drug and comparing their subsequent status with that of an untreated or a differently dosed group. When we wish to establish the precise effect of any experimental condition (let us keep leucotomy in mind as a concrete instance), we get right into the complexities of experimental design. All the usual means of demonstrating significant differences become relevant, notably *t*-tests for comparing means or chi-square for comparing frequencies of scores.

on tests pre- and post-operatively. Essentially similar in formal structure are problems of differences between various clinical groups with respect to experimental criteria.

How do investigations of this kind differ from the agricultural examples that lend Fisher's books what bucolic flavor they have? When different fertilizers are applied, or different strains of wheat used, there may be many variables which need to be controlled, just as in the psychiatric model—soil constituents, rainfall, exposure to the sun, in place of length of hospitalization, premorbid intellectual level, or presence of organic illness. But how much easier it is to measure the yield of a patch of wheat than the degree of recovery in a group of leucotomized schizophrenics! Of course, one can be concerned about the wheat's height and the volume as well as the weight of the harvest. The *patterning* of these variables, however, is not likely to be considered significant. In the psychological experiment, by contrast, no one test score or clinical rating alone is crucial, nor can it stand for the total effect, the change in the configuration of such data must be analyzed. But how does one test statistically the significance of a difference in configuration?

A *third* kind of clinical research goes a step beyond the establishment of differences, and tries to discover functional relationships. Variables are assumed to exist on both sides of the equation; the researcher tries to observe what happens to his criterion (let us say the Rorschach test) as successive increases in a human characteristic (such as anxiety) are studied. In such a problem the psychologist thinks first of correlation, though ideally one should work out the form of the relationship and fit a curve of some kind to it also. But, again if the true nature of clinical research is to be respected, the problem cannot often be so easily disposed of. Typically, the covariates will be a syndrome on the side of the patient, and a pattern of test results on the side of the criterion. What tools do we have for this kind of problem? Multiple and partial correlation? All very well if we are interested in only one variable on one or both sides, and if no discontinuities appear, and patterning is unimportant. Better not to have too many variables, however, because the

statistical labor gets very demanding and the formulas most complex, while if relationships are curvilinear, the same objections hold most forcibly.

All of this talk about syndromes and patterns points to one cardinal difference between clinical and classical experimental research. *It is taken for granted from the beginning in clinical research that the significant data are found in meaningful patterns, and that this patterning must be respected.* Whether the clinical psychologist or psychiatrist is working with a brain-injured patient or a normal personality, he has to consider the condition of his subjects as organisms, as people, and he has to keep this totality in mind all along.

Two important consequences follow. First, the nature of controls has to be quite different from that in the classical experiment; second, the techniques of analysis must enable the experimenter to deal with a shifting configuration of many variables or parameters, which may undergo "phase changes" or the emergence of new patterns in a discontinuous fashion.

Problems of Controls in Clinical Research

It is generally accepted that investigative science is essentially a matter of making observations under conditions over which the investigator has sufficient control to see causal relations clearly. Classically, this control has meant the holding of all important variables in a situation constant except one, the experimental variable, and then recording the effects of varying it on the criterion being measured or otherwise accurately observed. Such a design makes for clear, crucial experiments, it is easy to treat them statistically or even to derive an empirical function from the results. Thus, if we are interested in the properties of springs, we may first study the effect of various weights in stretching a spring, holding constant the nature of the spring used, the temperature, magnetic fields, etc. Having established the empirical equation, we may then hold constant the weight used to stretch the spring, and vary temperature, observing accurately the extension caused by, say, 50 grams, at 10°, 20°, 30° C. and so on—leading to another equation.

A recent trend in psychological research is the use of factorial

design. Recognizing the facts of interaction between psychological variables, it enables the experimenter to measure them by means of designs in which there is controlled change in more than one variable. It is a highly efficient type of design, permitting as it does the investigation of more than one possible cause of an effect in which we may be interested, as well as interactions between causes, where applicable it is to be preferred over the classical univariate design. It still requires more by way of direct manipulative control than the clinical researcher often has at his disposal, however. One must be able to arrange things so that given values of two or more variables or conditions will be attained simultaneously. It also has the disadvantage of limiting our attention to a single numerical score or index as a unit of analysis, out of several possibilities.

In much clinical research, not only can we not hold all relevant conditions constant except one, we must accept whatever variations occur, powerless to arrange them neatly beforehand. Since we cannot simplify our task by the usual means of control, we seek the control that is given by as exact knowledge as possible of the values of the uncontrolled variables, as we find them. In studying the effects of ego-involvement on levels of aspiration, the clinical researcher knows that he cannot insure that the many facts of personal history that may affect the criterion, statements about goals, will be the same for all of his subjects. Consequently, he tries to make the best of a bad situation and find out as much as possible about the people who are his subjects. He finds himself dealing with a complicated, if not tangled, web of inter-related factors, particularly if he chooses to observe more than one criterion aspect of behavior. And it does tend to be characteristic of clinical research to woo complexity of this kind too.

Of course, the fact that the subjects of study are human beings, often sick ones who are looking for help, dictates that human considerations must come before logical and mathematical ones. We cannot manipulate people in important ways just to help along the nicety of our controls. Furthermore, very often when we try to do so—when we create artificially "simple" situations, for example, where we may fancy that control of the classical kind has been attained—we find that we are

dealing with *different* subjects, who are reacting *artificially* to an *artificial* set-up. When we try to experiment in this way on any of the really important aspects of human emotional life, we are likely to create an awkward and false atmosphere. That is part of the meaning of the statement that existing configurations have to be respected in clinical research.

Problems of the Statistical Analysis of Complex Data

The clear consequence is that much more subtle and devious statistical methods than usual are needed if we are to try to unravel the causal nexus, to remove the effects of inescapable confounding of variables. It is out of the question to report clinical research of this character by the formula for an empirical function. Most of the techniques I am familiar with, such as multiple and partial correlation, analysis of variance and covariance, seem to offer promise and yet to bring up serious difficulties.

A Difficulties Attending the Use of Familiar Techniques

To begin with, (1) the numbers of subjects are inevitably limited in any research that seeks to gauge the important aspects of each subject's personality. With few degrees of freedom to work with, the more complicated correlational methods become unusable. The attempt to get larger N's through lumping together essentially different groups, or skimping on the thoroughness of study, destroys precision instead of increasing it. (It is true, the trend seems to be toward cooperative clinical research. Many hands can often get together sufficient data on rather large numbers of cases, for many kinds of clinical research, if personnel and money are no problem.) (2) Problems of non-linearity often vex correlational analysis. (3) Turning to the analysis of variance, the experimenter is all too likely to find that his data are not homoscedastic—variance is not uniform enough throughout the tables of results for as complex an analysis as he wants.³ The end result may be that the researcher whose statistical sophistication extends no further than mine

³ During the round table discussion, Dr. Phillip Kilon suggested in reference to this point that a transformation of the data (as by converting them into logarithms) often overcomes their heteroscedasticity and makes the analysis of variance applicable without affecting P values.

may find himself using simple methods anyway, willy-nilly. At the same time he will recognize that he is losing much of the richness of his data, but not know what to do about it. What is worse, he may feel impotently aware that many of the differences or relationships that he obtains may be exaggerated or underplayed because of the confounding effect of other variables, the effects of which he was not able to remove or hold constant.

Not all clinical research fits this model, of course. Some of it deals with such complexities that measurement of any kind seems futile or impossible, for example, psychiatric research on the dynamics of a neurosis. At other times it more closely approximates the classical experimental model. But the hallmark of clinical research, generally speaking, is that it tries to deal with complex subjects in complex, more or less natural settings.⁴ The general statistical problem of most clinical research, then, is roughly the same: how to deal with highly patterned, interrelated data.

Let me give some more specific examples of this kind of statistical problem. I have just been working with the problem of quantifying self-insight. Even though I had self-ratings and criterion-ratings on a large number of personological variables, and though something about their patterning on each subject was available from case studies, I was forced to use a summative atomistic measure of insight. Aggression might be the most crucial problem for this man, and relatively unimportant for that one, but it had equal weight with other needs, some of which might be quite trivial in the lives of both. Should a system of differential weights have been used? It would have been clumsy and approximate and would in no way have expressed

⁴ Here some note has to be taken of the many thorny problems contained in the easy words *simple* and *complex*. As a rough first approximation, let me offer these considerations: (1) Something is *simple* if it requires a few concepts or few coordinates to describe its structure, *complex* if it requires many. (2) An event is *simple* if it can be explained to a certain margin of error by a few determinants, *complex* if you need to isolate many determinants in order to reach the same precision. A further significance assumed here to be implied by complexity is the number of relationships between the parts of a whole, and their degree of order (hierarchy, symmetry). I have not intended to imply that there is any constant relationship between the concepts *simple-complex* and *artificial-natural*. It should be apparent, also, that simplicity is always relative to one's purpose and approach: a rose may be simpler than a symphony to an artist, more complex to a physicist.

the patterning of the needs in each man—their fusions, conflicts, subordi-
nations or hierarchical relationships.

Consider a simpler problem—expressing the degree of match-
ing or agreement between two series of simple patterns. A
somatotype is a simple pattern of three numbers, each number
representing the degree to which one of three components of
physique is present. Two somatotypers make independent rat-
ings of a group of subjects, and wish to find out how reliable
their judgments are. They can, of course, correlate each com-
ponent separately, and if the correlations are all plus one, the
patterns must be in agreement (excluding constant bias). With
lower but still quite respectable separate reliabilities for the three
components, there can be quite a lot of important disagree-
ment over which component is dominant in any one physique.
There is a good deal more difference between a 2-4-5 and a 2-5-4
than there is between 2-4-5 and a 2-3-4; but how can such
very simple pattern differences be handled statistically?

When we deal with a much larger array of numbers, such as
are found on the formal psychogram or summary sheet for a
Rorschach test, or the subtest scores of the Wechsler-Bellevue
scale, how much more hopeless it seems to try to relate such
patterns to anything, mathematically! Yet one important kind
of clinical psychological research, the validation of tests, has
to approach these data from a configurational point of view.
Consider one of the simplest problems in Rorschach validation,
since a single criterion, IQ, is involved: the estimation of intel-
lectual level. The books tell us that the number of whole re-
sponses, especially well-organized ones, the number of move-
ments and their quality, the accuracy of form perception, and
the number of good original responses (among other things)
are positively related to intelligence. But the patterning is such
a crucial matter that, as Klopfer remarks, only the $F+$ per-
cent gives an appreciable correlation when these factors are
correlated with IQ. The most subtle application of multiple-
regression methods would not help much, either, since there
are so many other things than intelligence which can affect any

¹ Dr. Rulon said in the course of the discussion that multiple correlation (Rin *us*)
would handle this problem adequately, though it seems that, even so, one must assume
no constant differences between raters.

of these variables. Yet judgments of skilled analysts of the test, taking the above and many other aspects of the record into account in their interrelations and mutual implications, agree fairly well with intelligence tests.

B Special Difficulties in Validating Psychological Tests

When we turn to the validation of diagnostic conclusions from any one test, or especially a battery, the statistical problems pop up from behind every test score.

There are a number of important problems that may properly be called *pre-statistical*. First of all, there is the group of very vexing problems of reducing clinical data to quantitative form in which they can be handled by statistics. Few research procedures in the field, even including tests, result directly in meaningful numerical indices. Ratings are often resorted to of necessity, with all the headaches that they bring. The principal statistical problem involved is getting numbers which mean something more than a denotative or ordinal scale.

Second, it must be recognized that test patterns do not indicate directly the presence of particular psychiatric diseases. Rather, they reflect discrete aspects of personality, of intellectual functioning, of thought organization. Therefore, they must be validated against these aspects rather than against a notoriously unreliable nosology. If, on the contrary, we were to follow the advice of those who urge the validation of Bellevue scatter analysis through multiple correlation, not only would our diagnoses be bare statements rather than pictures of personality under the effects of illness, but we should be tied to the nosology on which the original validation study was done. There has been for some time a current of growing dissatisfaction in psychiatry with the standard nosology; psychologists would be setting their faces toward the past if they ignored this trend. We must try, then, to learn the relationships between our test data and the elements of mental disorder (such as anxiety, projection, psycho-motor retardation) which are variously patterned in different nosologies. We need differently designed validation studies, which will tax more severely the kind of clinical collaborator for whom diagnosis is only label-giving, but which will be easier for the good clinician, who is usually

surer of his *observations* of such phenomena as tension and obsessive thinking than he is of the diagnostic pigeonhole into which he has to cram them.

In the third place, there is the fact that validation studies, if done only by the use of clear-cut cases, do not relieve the clinician of the need for considerable ingenuity, and artistry if you will, when he comes to grips with the mixed picture that the everyday case presents. He may be able to find in the book the Bellevue scatter patterns to be expected for intra-cranial pathology, for hysteria, and for schizophrenia. But how is he going to be able to use them in the diagnosis of a schizophrenic process developing in a person of hysterical character make-up after trauma to the central nervous system? Variations in the test scores of most patients can be attributed both to present illness and to pre-existing modes of adjustment. When patterns are imposed upon patterns in these ways, can one not be forgiven for despairing of the possibility of diagnosis through multiple regression, or even of the possibility of ever fully validating everyday diagnostic use of tests?

The fact is that that marvelous unconscious statistician, the human brain, *can* learn to separate such overlapping patterns and make sense out of them. It is for this reason that the clinical psychologist unabashedly relies on what experience teaches far more than on what can be demonstrated statistically to the satisfaction of meticulous methodologists. "Experience" can take into account the crucially important qualitative analysis, what Rapaport calls "the tune of the record," which, by its very nature, can hardly ever be treated mathematically. It can take advantage of unconscious learning, the effects of subliminal recognition of cues the nature of which the clinician may be unaware. Even though that kind of so-called intuitive or artistic diagnosis may not directly contribute to science, and though it is difficult to pass on to others, *its successes must occur before they can be subjected to systematic study and their bases finally discovered*. Perhaps, then, the experienced clinician may be forgiven for an attitude toward statistics which often approaches the condescending. He knows that the methods of diagnosis his experience has taught him are under a constant validating check, their agreement with clinical data. He knows

also that statistical researchers can give scientific respectability to some of this knowledge, bit by bit, following at a considerable distance behind, but he has yet to hear of any contribution to diagnosis that was made by statistical research before it had been discovered by the working clinician

But let us suppose that these prestatistical hurdles have been successfully leaped, and a study is under way to validate certain supposed test indications of schizophrenia. Let us suppose, further, that the cases have been divided clinically in the way just recommended according to aspects of the disintegration of control over thought and emotion, for example. We come right up against the fact that different indicators mean particular aspects of schizophrenia in different cases. In one patient, the chaotic response to and use of color in the Rorschach test may indicate abandonment of a large degree of emotional control, while the TAT is relatively stereotyped, in another case, the Rorschach may be quite constricted, giving no hints about the emotional status underneath the surface inhibition, while wildly aggressive and sexual fantasies in the TAT indicate again the deterioration of control over emotion. Furthermore, one cannot reason diagnostically from the *lack* of a sign beyond a limited extent, just because a wealth of distant word associations indicates schizophrenic disorganization, an orderly association test does not necessarily prove the lack of such disorganization.

Why are these statements true? Certainly not because determinism in mental life is lacking or capricious. Rather, it seems to be that the structure of human organism, particularly of its psychic aspect, is an exceedingly complex matter of checks and balances. If A gives, but B and C hold, then no matter, again, no matter how strongly B and C stand fast, if D goes, then all is lost. We know in studying patients' histories, that a particular trauma, such as the accidental death of a parent in front of one's eyes, gives every sign of being directly pathogenic in one case, while this and other traumas may be piled on another, constitutionally strong person, or perhaps one with a good infancy, and only a slight degree of pathology occurs.

It seems to be doubtful, then, that in clinical research in etiology or anything else psychological, Koch's postulates can

be "the tablets of the law," as Dr. Kubie said in the 1946 Orthopsychiatric Round Table on Clinical Research.⁶ These postulates, you will remember, are "1) we will have to find x , the suspected cause in every patient suffering from a specific disease; 2) this x must be found in such patients only; and finally, 3) the experimental introduction of x into the host must produce the disease." Dr. Kubie believes that the first two principles are applicable to mental disease, even though determinants of human behavior are manifold, and though it is difficult to distinguish in psychiatry between what is the essential content of a disease and what is saprophytic or secondary content. Large numbers of cases will clear these matters up, he believes. But, consider the well-known hypothesis of Freud, that homosexual conflicts are a specific etiological element in paranoid projections. Grant for a moment that it is possible to satisfy the first requirement, and that this suspected cause can be found in every paranoid case (although there are many clinicians who will not grant its universality). It still does not follow that homosexual conflicts are found in such cases only, or that where such a conflict is induced, the paranoid symptoms always occur. The issue cannot be solved by objecting that it has to be just one particular kind of homosexual problem; almost all degrees of acceptance and awareness of these forbidden impulses may be found in paranoid patients at the time the symptoms develop.

It seems that we must reformulate the principles laid down by Koch in the light of modern conceptions such as the one championed by Bellak in his recent survey of etiological theories of schizophrenia.⁷ He maintains that schizophrenia is a reaction type, a kind of syndrome which may be brought about by a mixture of organic and psychological determinants in any proportions, from the purely psychogenic to the purely somatogenic. Certainly, there must be a similarity between this and the basic position taken by the diagnostic tester, that any of a known but very wide range of test patterns can be indicative of a particular disease.

⁶ Brenman, M. (Chairman) *et al*. "Problems in Clinical Research." *American Journal of Orthopsychiatry*, XVII (1941), 196-230.

⁷ Bellak, Leopold. *Dementia Praecox: the Past Decade's Work and Present Status; a Review and Evaluation*. New York. Grune and Stratton, 1948. Pp. xv, 456.

The statistical implications are obvious enough. In validation research, it may be extremely difficult to get enough cases showing any one test pattern to establish its relationship to the clinical criterion. Contaminations in the Rorschach test are widely accepted as pathognomonic of schizophrenia, for example, yet if one wanted to validate this statement by a statistical study, it might be necessary to test hundreds of patients before a half dozen clear-cut contaminations were found.

The considerations that have just been mentioned all point to the need for statistical techniques which, if not new, are at least unfamiliar to most clinical researchers. Perhaps we are asking for the impossible when we say that we would like means to handle the simultaneous variation of many variables, sometimes curvilinear, sometimes discontinuous, and relationships between patterns, all with relatively small numbers of cases and preferably without the necessity of too much computation, please. If we cannot have the moon, and if the statisticians cannot show us why we do not really need to have it, what does it seem to us that we must do?

Some Suggested Solutions

A. One direction for research to take has already been mentioned. That is toward less emphasis on quantification and statistics and more on careful observation and the attempt to understand what one sees. Rather than continuing to apply obviously inadequate statistical methods, clinical researchers might do much better to concentrate on intensive studies of single cases, observing in as controlled a way as possible, trying to discern meaningful relationships and to set up hypotheses which may be tested when appropriate methods for establishing proof are at hand. It is all too often forgotten that statistical methods are primarily ways of proving (or more exactly, disproving) hypotheses and only secondarily means of finding something out. The method of Freud and the other great clinicians who have contributed the most of our knowledge about the kinds of problems dealt with in the field we are discussing, was the method of discerning. Köhler offers convincing arguments that causal relations may be as directly perceived as

anything else, and organismic methodologists have taught us that the single case is and must be lawful. As a method of proof it leaves a good deal to be desired, but it is certainly a legitimate method of clinical research.

So great is the scientific glamor of more experimental methods that there is a danger that too little research of this phenomenological kind will be done. Our clinicians-in-training may come to think that a couple of good courses in methodology and statistics under their belts are a substitute for the necessity to keep as sharp, unbiased and fresh an eye as possible on the patient or subject himself. The aim of observational research is not only to find uniformities and pathognomonic signs. Just as much, it must strive to look out for the exceptions, for the unexpected and unexplained deviation from what textbooks and previous experience have told us. Great strides in clinical psychology and psychiatry will still be made through the discovery of new puzzles, new effects or phenomena to be explained as well as through better means of handling familiar types of data.

B. It is not usually necessary to caution clinicians against plunging too abruptly into quantitative treatment of their data, but there is undoubtedly a good deal of research in the field that suffers for this reason. As soon as a method yields quantitative data, there is a strong temptation to subject them at once to statistical analysis. Actually, one can waste a lot of time in this kind of thing if he has not made sure first that he has chosen the *appropriate level of abstraction* on which to do the analysis. Dr. Horn's experiment in the diagnostic process is, I think, a good example of the *second direction* for research to take: statistical analysis on the proper level of abstraction instead of treating the most obvious results.⁸ Rather than trying to relate the quantitative results of the Rorschach, TAT and other tests directly to characteristics of personality, he had his judges study each test and, using it as best he could, make ratings of these characteristics.⁹ Thus, the complex patterning

⁸ Horn, Daniel. "An Experimental Study of the Diagnostic Process in the Clinical Investigation of Personality." Harvard University, 1953. Unpublished Ph.D. thesis.

⁹ After hearing the Round Table, Dr. Lee Cronbach kindly reminded me of another important type of design which uses data on this level of generalization, but without

remained in the mind of the clinician, who could use the most intricate interplay of reasoning from qualitative and quantitative evidences in order to arrive at his judgments. By using enough judges, some especially skilled in one method or test, others in another, he could arrive at a meaningful assay of the usefulness of each test. The important research on the selection of clinical psychologists under the direction of Dr. Lowell Kelly is another good example of this approach, on a much larger scale. Premature quantification, on the other hand, can produce the kind of sterility that is seen in much so-called "objective" social psychology and sociology, where the easy availability of certain superficial kinds of quantitative data has raised false hopes of immediately discovering mathematical laws of social behavior.

C. A *third direction* clinical research may take has been suggested by a recent clinical study of hypnotizability, by Roy Schafer.¹⁰ With a sizeable and quite mixed group of patients, he could find no single test score or indicator which had a reliable relation to hypnotizability. When he made blind analyses of the battery of tests that had been given each subject and wrote out careful and complete descriptions of their personalities, a number of clear and statistically reliable differences between the good and bad hypnotic subjects appeared in terms of kinds of ego structure and the like. This is another example of the second direction, quantification on the proper level, which shows, incidentally, how statistics may be used in combination with the case-study method. Here, however, is the main point. Recognizing the necessity of applying his newly discovered criteria to another group (for "cross-validation"), Schafer repeated the procedure, this time using as subjects a group of doctors, all candidates for advanced professional training in psychiatry. In this quite homogeneous group, not only were the previous findings sustained, but it was now possible to find significant differences in particular test scores between good

the necessity of ratings, matching. He has since published an account of his own, promising extension and sharpening of the method of matching, in his article, "A Validation Design for Qualitative Studies of Personality," *Journal of Consulting Psychology*, XII (1948), 365-374.

¹⁰ A summary of this research was presented by Dr. Schafer at the APA meetings in 1948 (Abstract in *American Psychologist*, III (1948), 280.)

and bad subjects. This very striking finding takes us back to the analysis of the kinds of controls that are possible in clinical research. In the heterogeneous group, the simplicity of relationship between an experimental variable, such as a test score, and the criterion (hypnotizability) was obscured by the many uncontrolled sources of variation in each. One kind of control -- the usual one in clinical work -- was the psychologist's understanding of the significance of these sources of error through the total pattern of test results, so that he could transcend them by working at a higher level of integration. By working with a homogeneous population of subjects, the other kind of control was attained, many sources of error were held constant. Thereby many former *variables*, which could not have been held constant statistically, became *parameters*, and simple relationships on a lower level of analysis became apparent.¹¹ The patterning was still there, but it was in part implicit, so to speak, in those parameters. Whether the test correlates of hypnotizability would still appear at another value of the parameters could only be determined by further experiment. If they did not, we should have a good example of the kind of discontinuity and emergence that is met with in clinical data.

What are the implications of these findings for clinical research? They offer the hope that the need for complex statistics may be obviated in many cases when the researcher can work with highly homogeneous populations. Rather than try to find "typical" Bellevue scatter patterns for each of a variety of nosological groups, research on scatter (or any other test patterns) might proceed in the following manner. First, let us study a sizable group of obsessive neurotics, all of whom come from comparable socio-economic backgrounds. Let us take some of the outstanding aspects of obsessive neurosis, such as anxiety, intellectualization, ruminativeness, etc., and obtain quantitative ratings of each symptom for each patient. Then we may be able to find clear relationships between each constituent and simple scatter patterns, such as the discrepancy between In-

¹¹ The sense in which this pair of terms is used here may be clarified by reference to the experiment with the springs, above. When different weights are used at a constant temperature, weight is variable, temperature a parameter. Conversely, when the spring's extension is studied by *varying* temperature, a *parameter* of the obtained function is the value of the constant weight used.

formation and Vocabulary subtest scores. Then we shall turn to a similar group of hysterics, or perhaps to another group of obsessives but at a very different cultural level, and so on down the line. In this way we should know our parameters, and might be able to validate much in scatter analysis that escapes the methods that have been tried so far. Other applications of this general principle need not be spelled out here. It is perhaps worth while to mention, however, that it should by no means be restricted to dealing with nosological entities.

One further advantage of this proposed technique of research is that it would tend to reduce the complexity of findings. As it is, the conscientious, scientifically scrupulous man in clinical research is caught in a distressing dilemma. On the one hand, if he respects the subtlety and intricacy of human functioning and tries to show it in his work by measuring all relevant variables, his study gets more and more nearly impossible to carry out, impossible to analyze, and the resulting report impossible to get published or read. On the other hand, if he oversimplifies so that the dimensions of his study are comfortable to handle in these three respects, he knows that he may be producing a caricature of life. Perhaps this method of homogeneous groups, together with the method of working on a high level of integration, may provide a way out.

As in the physical example several times referred to, we had to repeat the experiment a number of times with each of several kinds of springs, just so in clinical research we shall have to abandon the false hope of solving a problem by a single study on a single group. Perhaps the necessity for repetition on different kinds of groups might bring about that end of experimental isolation and beginning of cooperation recently called for by Fiske in the *American Psychologist*¹²

How Statisticians Can Help

May I conclude with one last request of the statisticians? Throughout this paper, when I have asked for the development of new methods, I have been plagued with the uneasy feeling that many of them probably existed and that I would know

¹² Fiske, D. W. "Must Psychologists be Experimental Isolationists?" *American Psychologist*, I (1947), 23-28.

about them if I read *Psychometrika* and the other statistical journals. Another source of my request comes from the realization that results all too often depend on the statistical method used, and that many clinical researchers' ignorance of a wide variety of methods makes them unnecessarily rigid and limited. There are certainly many books on statistics already available, but so far I have not found the one that I want, and I wonder if anyone has tried to write it.

It would not be a text, but a handbook of statistical methods for clinical and other psychological research workers. It would be a systematic compilation of formulas and methods, grouped under the usual headings (Measures of Dispersion, of Correlation, etc.). With each formula, there would be the following information:

- (1) The assumptions underlying its use, with concrete examples of what they mean in terms of psychological data.
- (2) Some indications of the degree to which each assumption can be violated with impunity, and what the results of violations will be.
- (3) A discussion of the principles determining the number of cases needed.
- (4) Some examples of the kinds of problems to which the technique is appropriate, with perhaps some examples telling when it is inappropriate.
- (5) The quickest method of computation, both with and without the use of calculating machines, possibly also indications for applications with IBM tabulators or other machines.
- (6) I should particularly like to see expositions of short-cuts for the calculation of direct probability.

I recently learned by communication from a mathematical statistician¹² something I was unable to find in any statistics book I could lay my hand on: a method for getting summations of binomial expansions. The binomial theorem is an exact method that can often be used instead of the inexact approximation of Chi-square, especially with very small samples, but direct computation and summing of terms are very laborious. But, as I learned, one can enter the tables of the incomplete

¹² Dr. Albert H. Nowker, whom I wish to thank for his valuable up-

beta-function¹⁶ with a minimum of effort and read off these summations exact P-values for the deviation of any obtained figure from a hypothesis. Perhaps there are other wrinkles of this kind which should be brought to light.

The book I am hoping for will accept the fact that most of the people who use it will not be very good mathematicians and would get little enlightenment from a derivation. In its demands on mathematical and statistical sophistication, and in its lucidity of exposition, it will be as much an opposite to Fisher's "Statistical Methods for Research Workers" as possible. I see it as necessarily a product of collaboration between mathematical statisticians and other men (including if possible, a clinical researcher) whose primary jobs are in research but who know statistics well enough to communicate easily with them. I cannot guarantee the sale of more than one copy, but I can guarantee that mine would be well-thumbed.

¹⁶Pearson, Karl (1914) *Tables of the Incomplete Beta function*. London: University of London Press, 1914. Pp. 165.

THE OPINIONS OF SYRACUSE UNIVERSITY STUDENTS ON SOME WIDELY DISCUSSED CURRENT ISSUES

N. M. DOWNH

State College of Washington

and

C. H. PACE and M. J. TROYER

Syracuse University

DURING the academic year, 1947-1948, Syracuse University carried on an all-university self survey which would furnish information for intelligent planning during the years ahead. Included among the various concerns of the survey was an investigation of the program of general education of the University.

As a part of this study of general education, a sampling of seniors, Class of 1948, and of sophomores, Class of 1950, was asked to respond to a battery of opinion scales made up of statements on issues that are the subject of rather wide discussion at the present time. This battery, which is reproduced below in Table 1, consisted of ten scales covering the following topics: politics, government, civic relations, the world, experts, science, philosophy, music, art, and literature. The students were asked to check whether they agreed, had no opinion at all, or disagreed with each statement.

Perhaps, opinions and attitudes cannot be considered right or wrong. However, if one of the objectives of general education is to develop attitudes and opinions in the students which are consistent with democratic values and which lead to effective participation in our democratic society, then we will want to know how well these opinions have developed in the students. Also, if there are opinions about various topics which most specialists in that field hold among themselves, then we can consider students' opinions on these topics to be desirable or undesirable from the point of view of whether they agree or

TABLE 1
Responses of Students and Faculty on the Opinion Scale (Percentages)

				POIITICS	Faculty N 30 Students N 555
1	2	3			
F 10	7	84	1	Sending letters and telegrams to congressmen has little influence on legislation.	
S 29	4	67			
F 3	3	93	2	Political parties are run by insiders who are not concerned with public opinion	
S 23	5	72			
F 29	4	67	3	When the public is really concerned about an issue, its judgment is usually correct and unassailable, no matter how complex the issue	
S 13	5	82			
F 17	0	84	4	In some elections there is not much point in voting because the outcome is fairly certain	
S 16	2	82			
F 70	3	26	5	Pressure groups are useful and important features of representative government	
S 56	8	36			
F 73	3	24	6	On most issues we should expect our representatives to vote according to their convictions, even though they may not always reflect the opinion of their constituents	
S 57	6	57			
1	2	3		GOVERNMENT	Faculty N 30 Students N 555
F 15	0	86	7	The best government is one which governs least.	
S 25	7	68			
F 23	7	70	8	Democracy depends fundamentally on the existence of free business enterprise.	
S 62	6	32			
F 35	3	62	9	Communism and Fascism are basically and historically similar.	
S 39	13	48			
F 13	7	80	10	The most serious danger to democracy in this country comes from Communists and Communist-dominated organizations	
S 46	6	48			
F 13	3	83	11	Government planning should be strictly limited, for it almost inevitably results in the loss of essential liberties and freedom	
S 23	13	64			
F 10	3	86	12	Individual liberty and justice under law are not possible in Socialist countries.	
S 27	15	58			
1	2	3		CIVIC RELATIONS	Faculty N 30 Students N 555
F 97	0	3	13	All Americans—Negroes, Jews, the foreign born, and others—should have equal opportunity in social, economic, and political affairs.	
S 88	3	9			
F 10	23	68	14	Familiarity breeds contempt	
S 17	11	72			
F 10	16	75	15	Foreigners usually have peculiar and annoying habits.	
S 11	15	74			
F 0	0	100	16	Children of minority groups or other races should play among themselves	
S 1	3	96			
F 60	10	30	17	Most children, these days, need more discipline.	
S 63	12	25			
F 10	14	76	18	Agitators and trouble makers are more likely to be foreign born than native Americans	
S 15	10	75			
1 Agree				F—Faculty	
2 No opinion at all				S—Students	
3 Disagree					

TABLE 1 Continued

				THE WORLD	Faculty N 39 Students N 555
1	2	3			
F	28	11	6	19 We are not likely to have lasting peace until the U S and the Allies are stronger than all the other countries	
S	40	5	55		
F	10		100	20 If we lower our tariffs to permit more foreign goods in this country, we will lower our standard of living	
S	27	17	56		
F	18	14	68	21 Deep ideological differences between countries are irremediable	
S	19	16	65		
F	3		97	22 If we allow more immigrants into this country, we will lower our standards of culture	
S	17	15	68		
F	25	4	71	23 The United Nations should have the right to make recommendations which would bind members to a course of action	
S	50	7	57		
F	82	1	16	24 Over the next decade, we must try to make the standard of living in the rest of the world rise more rapidly than in our own country	
S	54	15	31		
				EXPERIENCE	Students N 555 Not rated by faculty
1	2	3			
S	20	10	70	25 The predictions of economists about the future are no better than guesses	
S	5	6	89	26 Doctors' diagnoses of illnesses turn out to be wrong almost as often as right	
S	18	7	75	27 Parents know as much about how to teach children as public school teachers know	
S	5	4	91	28 The findings of psychologists are not helpful in fitting workers to jobs	
S	86	7	7	29 Contemporary painters, designers, playwrights, and musicians are engaged in work as important as my own	
S	91	2	5	30 A person in a skilled trade is worth as much to society as one in a profession	
				SCIENCE	Students N 555 Not rated by faculty
1	2	3			
S	85	7	8	31 There are many worthwhile and important concepts which can not be proved scientifically	
S	81	8	11	32 The harnessing of atomic energy will bring about fundamental changes in our economic and social order.	
S	80	10	10	33 The government should promote and subsidize research in the social sciences	
S	78	10	12	34 There will be as many or more scientific discoveries, inventions, and technological changes in the world during the next fifty years as there were during the past fifty years	
S	65	5	30	35 We now have enough scientific and technological knowledge to substantially eliminate poverty, disease, and ignorance in the world, if we would only apply our knowledge	
S	84	8	9	36 The government should promote and subsidize research in the physical and biological sciences	

1 Agree
2 No opinion at all
3 Disagree

F—Faculty
S—Students

TABLE 1 -Continued

			PHILOSOPHY	Faculty N 10 Students N-555
1	2	3		
F	0	0	100	37 What one does with his life is not very important, except to oneself
S	8	2	90	
F	0	0	100	38 If the goal is worthwhile, almost any method is justified in attaining it
S	14	2	84	
F	80	0	20	39 Personal integrity of conduct and continuous searching for truth are the most important goals in life for me.
S	54	20	26	
F	100	0	0	40 A contract is morally binding, one should never default on his pledged word
S	75	7	18	
F	0	0	100	41 Religion has little to offer intelligent, scientific people today
S	11	7	82	
F	10	0	90	42 The greatest satisfactions in life for me come from financial success, influence, and prestige
S	15	8	77	
			MUSIC	Faculty N 14 Students N 555
1	2	3		
F	7	7	86	43 What is good and bad in music is a matter of personal taste
S	75	4	21	
F	0	0	100	44 The tendency of some modern composers to use strange harmonies and discords makes poor music
S	10	15	75	
F	0	0	100	45 Music is a form of expression which normal people are not capable of understanding.
S	5	4	91	
F	7	0	93	46 The main thing about good music is its lovely melodies
S	27	14	59	
F	0	0	100	47 There has been little or no outstanding music composed in the 20th century
S	7	9	84	
F	100	0	0	48 Radio should give people much more opportunity to hear good serious music
S	83	9	8	
			ART	Faculty N- 20 Students N 555
1	2	3		
F	5	5	90	49 What is good and bad in art is a matter of personal taste
S	71	4	25	
F	0	0	100	50 Art is a side line, not part of the main business of life
S	14	7	79	
F	0	0	100	51 Modern painting—impressionism, expressionism, cubism, surrealism and the rest—is mostly the work of crack-pots
S	16	13	71	
F	95	5	0	52 Many paintings which were considered radical at the first are regarded as classics today
S	80	15	5	
F	50	10	40	53 New houses should be built of modern design rather than Colonial, Cape Cod, Spanish or some older style
S	17	16	67	
F	75	15	10	54 Electric lighting fixtures should not look like candles
S	30	40	30	
1 Agree			F—Faculty	
2 No opinion at all			S—Students	
3 Disagree				

TABLE 1-Continued

				LITERATURE	Faculty N 20 Students N 555
1	2	3			
F S	6 8	9 18	100 74	55 Writers who use a different and cryptic style—such as Gertrude Stein or James Joyce—should not be taken seriously	
F S	95 76	5 15	0 49	56 Good writing by promising young writers is more likely to be found in Harpers or the Atlantic Monthly than in Colliers or The Saturday Evening Post	
F S	0 9	0 9	100 82	57 Literature should not question the basic moral concept of society	
F S	0 7	6 4	95 89	58 Hollywood versions of novels or plays are usually as good as the originals	
F S	50 56	0 16	10 28	59 The current list of best sellers in fiction and non fiction does not provide a very good index of literary merit.	
F S	20 31	10 20	70 49	60 Literature should be judged primarily by its contribution to our understanding of the social order	
1 Agree				F—Faculty	
2 No opinion at all				S—Students	
3 Disagree					

not with the individuals who have devoted a large amount of study to the particular field.

Different parts of the opinion scale were completed by faculty members teaching in the area with which each part was concerned. "Politics," "Government," "Civic Relations," and "The World" were submitted to all personnel of the Maxwell School of Citizenship of professorial rank and to a sampling of instructors. The part entitled "Philosophy" was answered by members of the Departments of Philosophy and Bible and Religion and the chapel staff "Music" was presented to all staff members in the School of Music and "Art" to the staff of the School of Art. The "Literature" section was given to selected members of the English Department who were most concerned with the teaching of literature. The parts labelled "Experts" and "Science" were completed by no particular group. Eighty per cent or more of the staff members who filled out these scales had been at Syracuse for at least two academic years.

The response for each item which the majority of the experts marked was taken as the correct response for that item. The students' papers then were scored, using these faculty responses as the key. Three items were not used because of insufficient agreement among the experts.

The students who participated in this part of the survey were selected from five colleges of the University Applied Science, Business Administration, Fine Arts, Home Economics, and Liberal Arts. Mean scores for each part of the scale and for the entire scale for both classes in these five colleges were computed. Homogeneity of variances was compared by the "F" test and the significance of the differences between the means by the "t" test.

When the students' responses were compared with those of the Faculty, a rather large difference of opinion appeared for some items (Table 1).

On item 6, concerned with government representatives voting according to their convictions, even though they are not reflecting the opinions of their constituents, 73 per cent of the Faculty agreed with the item and only 37 per cent of the students. Item 8, which stated that democracy depends fundamentally on the existence of free business enterprise, was marked "disagree" by 70 per cent of the staff members and 32 per cent of the students. "The most serious danger to democracy in this country comes from Communists and Communist-dominated organizations," item 10, was marked "disagree" by 80 per cent of the Faculty and by 48 per cent of the students. Eighty-six per cent of the staff and 58 per cent of the students disagreed with the statement that individual liberty and justice under law are not possible in socialist countries (item 12). Item 24, which stated that in the next decade we must try to make the standard of living in the rest of the world rise more rapidly than in our own country, was agreed to by 87 per cent of the faculty and 54 per cent of the students.

In the area of philosophy, three of the items showed differences between students and staff. Item 39, which stated that personal integrity of conduct and continuous searching for truth are the most important goals of life for me, was agreed to by 80 per cent of the staff and 54 per cent of the students. All of the Faculty believed that a contract is morally binding and that one should never default on his pledged word, item 40, as compared with 75 per cent of the students. Eleven per cent of the students believed that religion has little to offer intelligent, scientific people today and 7 per cent of the students had no

opinion at all on this statement (item 41) The Faculty rejected this viewpoint completely

In the area of the Fine Arts, item 43, which said "What is good and bad in music is a matter of personal taste," was marked "disagree" by 86 per cent of the staff and 21 per cent of the students. A similar item on art, item 49, was similarly marked by 90 per cent of the staff and 25 per cent of the students. In the area of Literature, item 56, which stated that good writing by promising young writers is more likely to be found in *Harpers* or *The Atlantic Monthly* than in *Colliers* or *The Saturday Evening Post*, was marked "agree" by 95 per cent of the staff members and 36 per cent of the students.

The responses of the Senior Class were then compared with those of the Sophomore Class. Chi square was used as a test of significance and the responses of the two classes were found to be significantly different for twelve items. These items were 1, 18, 25, 34, 41, 42, 44, 48, 49, 54, 55 and 56. Of these twelve items, nine of them showed the opinions of the sophomores to be more in agreement with the opinions of the Faculty. The exceptions to this were items 34, 48 and 54.

When scores on the scale were compared, no difference was found between the mean total score of the seniors and that of the sophomores. The Liberal Arts sophomores and the Fine Arts seniors were significantly above the all-university mean of their respective classes and both classes of the College of Applied Science and the Senior Class of the College of Business Administration were significantly below their respective all-university means.

On the part labelled "Politics" the Liberal Arts sophomores were significantly above the mean of all sophomores and the Fine Arts sophomores significantly below it. Both classes in all colleges were well below the mean for the Faculty. In "Government" both classes of the College of Liberal Arts were significantly above their respective class means, the sophomores in the College of Home Economics and both classes of the College of Fine Arts significantly below it, and all groups below the faculty mean. In "Civic Relations" the Liberal Arts seniors were significantly above and the Business Administration seniors were significantly below the all-university mean for this class. In this section the mean for the faculty and the

means for the two classes were approximately the same. In their opinions concerning "The World," the Liberal Arts sophomores were significantly above their class mean and the Home Economics sophomores were significantly below it. In this area the seniors more closely approximated the faculty mean than did the sophomores. In the area on "Experts" no significant differences were found between any of the college and university means. On the "Science" part of the opinionaire, the Liberal Arts seniors were significantly above the all-university mean for seniors.

In the part on "Philosophy" no differences were found and both classes were lower than the faculty mean. In "Music," the Fine Arts seniors and Home Economics sophomores were significantly above their class means and both classes of the College of Applied Science and the Business Administration seniors significantly below it. In "Art" both classes of the Colleges of Fine Arts and Home Economics were significantly above their respective means and both Applied Science classes and the Business Administration seniors significantly below them. On "Literature" the Fine Arts seniors were significantly above the all-university mean for this class and the Applied Science seniors significantly below it. On these last three sections of the opinion scale, the mean faculty opinion score was much higher than the student mean score.

When the two classes in the same college were compared, significant differences were found in a few cases—the Liberal Arts seniors were significantly higher in "Science" than the sophomores in the same college, the Business Administration sophomores higher than the seniors in "Civic Relations," and Fine Arts seniors higher in "Civic Relations," "The World," "Art," and "Literature," and in the College of Home Economics the seniors were significantly higher than the Sophomores in "The World," "Music," and "Art."

Summary

1. There is considerable difference of opinion among the students of Syracuse University on widely discussed current issues. On most of the items in these opinion scales, the majority of the students agreed with the Faculty.

2. On a number of the items, however, the replies of the students differ considerably from those of the experts.

3. Seniors generally reflect opinions which are no closer to those of the experts than are the opinions of the sophomores.

4. In general, students specializing in an area reflect better opinions in their own area. The farther one gets away from his own area the wider do his opinions differ from those of experts. Yet the statements of opinions included in the scales were designed to reflect basic insights and generalizations about various topics rather than points of view which would be peculiar to the specialist. The need for broader general education is therefore suggested.

THEORETICAL PROBLEMS IN THE SELECTION OF STUDENTS FOR PROFESSIONAL SCHOOLS

EDWARD J. FURST

University of Michigan

SELECTION is commonly carried out in professional schools as though it were a process independent of the educational program. Admissions tests are given, candidates ranked, and selections made with little or no reference to the conduct of other aspects of the school's work. In this process, tests lend a certain appearance of respectability as well as of objectivity, although their effectiveness may not have been empirically determined by the school. Possibly much of this difficulty stems from the fact that the tests are ordinarily developed by outside specialists and so are not accepted by the staff as integral parts of the total educational program. Whatever the causes of this difficulty, the fact remains that the planning and conduct of admissions procedures are not commonly integrated with the planning and conduct of other aspects of the school's program.

It is the purpose of the article to re-state the problem of selection in terms of a much broader framework than conventionally conceived. Most of the discussion will focus on three rather basic problems: determining criteria of proficiency, determining the predictive measures, and limitations to prediction. It will be assumed that prediction¹ is the essential element in the process of selection.

Determining the Criteria of Proficiency

Every selection program involves judgments as to the future performance of the candidates. Since future performance serves as a criterion against which selection procedures may be validated, it becomes necessary to clarify the nature of this per-

¹ Although the term "prediction" is used in this article, the writer recognizes that the most one can do is to make a statement of an individual's relative probabilities of success in specific activities rather than "predict" what he will actually do. Cf. Super (J. 654-663).

formance. Two broad kinds of performance may serve alternatively as the criterion we wish to predict—success in the professional school and success in professional work. In the terminology of Thorndike (3), the latter is the ultimate criterion, it is the ultimate criterion because it represents the final goal of professional education. In this sense the ultimate criterion is *the test* not only of the effectiveness of selection procedures but also of the program of professional education.

Success in Professional Work As stated, "success in professional work" is obviously too nebulous to be of use in guiding the validation of selection procedures and the curriculum of the professional school. The concept calls for considerable analysis, especially with reference to the definition of professional work and criteria of success in it.

It would be naive, of course, to talk of *the* position of doctor, lawyer, or teacher as though there were one uniform type. Professional positions vary from place to place and from time to time. Ultimately, there are as many positions as there are individuals engaged in a profession, for to a certain extent each position is determined by what the particular practitioner puts into it. This generalization is true of any occupation.

The technique of job description may be utilized as a means of classifying the major kinds of positions in a profession. Job descriptions of a wide sampling of positions indicate the activities performed by the practitioner, their relative importance, and the conditions under which they are performed. But the question of what *should be* the major types of positions in the profession still remains, for the descriptions do no more than describe and classify activities as they exist. One recent study, for example, found that about a quarter of a hospital nurse's time is spent in bathing and feeding patients and in clerical and routine duties, many of which might be performed as readily by non-nurses (1). This finding raises the question of whether or not nurses should be relieved of such duties so that they may devote themselves more completely to work of a professional nature. Doubtless this issue is very much alive today. The point is that the collection and classification of job descriptions simply provide us with normative data; they describe activities as they are, rather than as they ought to be. Illustrations can also be drawn from other professions. In teach-

ing, some individuals must devote so much time to the keeping of pupil records that they come to think of personnel work as synonymous with record-keeping. And in law, one wonders what the effect of procedures analysis and simplification studies, on the one hand, and the application of the personal frame of reference of modern psychology (2), on the other, would be upon the entire legal profession and the law schools. The major kinds of activities that make up a profession at any time are normative, at best.

Assuming stability of tasks and responsibilities which go into the major types of positions in a profession, job analysis, as distinct from job description, may be utilized to identify the characteristics needed by the worker for successful performance. It asks this question: "What kinds of knowledge, abilities, and traits does a practitioner need in order to be successful in a given type of position?" Answers to this question are of value to the professional school in setting up its curriculum. The kinds of knowledge, abilities, and traits required for particular kinds of professional activities suggest objectives for the curriculum. The objectives for the curriculum—that is, the kinds of learning to be fostered in students—give clues as to the particular aptitudes and traits required of the professional students.

Implicit in the process of job analysis is the idea that criteria of proficiency exist or can easily be established. Probably the day-to-day proficiency of the practitioner could be readily ascertained, if necessary. Ordinarily, however, indexes of proficiency covering a relatively long period of service are sought. Several such indexes or criteria have been suggested for any profession: length of service, salary or income derived from work, and rank or professional standing. The limitations of any one of these are easily inferred.

Criteria of proficiency are basic to any definition of "success in professional work," but they do not tell the whole story. Success goes beyond mere proficiency to take in the practitioner's satisfaction with his occupational pursuits, and so is subjective from that standpoint. Also, the public has a stake in the matter. Results, services rendered, are the chief touchstones of the public in appraising the success of a practitioner.

The discussion up to this point, in short, has emphasized

that success in professional work is an elusive and relative concept. Although the difficulties are great in defining the major types of positions in a given profession and in determining criteria of success, the basic fact remains that the ultimate test of admissions procedures and professional education lies in demonstrated success on the job.

Success in the Professional School.—This is a more immediate and direct criterion against which to validate selection procedures. Again, the meaning of success needs to be clarified—what should success include? Certainly it ought to include graduation from the school. This is, after all, the official act of the school indicating that the student is qualified to engage in the profession. But graduation also implies that the successful student was able to persist in the school, to complete the program of professional education despite the possible influence of certain factors which tend to cause withdrawal or elimination. This point needs emphasis because many unpredictable factors only remotely related to academic achievement can cause a student to drop out or be eliminated.

More specifically, of course, graduation implies that the student has achieved certain objectives of instruction. He has acquired a body of knowledge, developed certain skills and abilities, and cultivated certain traits which the faculty has judged necessary for him to possess as a prospective practitioner. The degree of his development in each of these major directions constitutes the several criteria which denote the degree of his success in the professional school. These are the kinds of achievement which selection procedures necessarily must attempt to predict, if there is to be any prediction worthy of the label "scientific."

These desired outcomes are crucial as guides to curriculum development, the selection of instructional materials and methods, the development of examinations and other evaluation instruments, and the development and validation of selection procedures. Hence it is desirable that they be carefully determined by the school. This is a curriculum problem, and well-established procedures for attacking it systematically have been developed by curriculum specialists. Such a systematic procedure, more elegantly termed a *rationale*, has been compre-

hensively outlined by Tyler (6). This procedure can be clarified by showing how it could be applied in a field—nursing education, for example.

A rationale for determining instructional objectives includes two basic steps, one concerned with the getting of suggestions as to possible goals and the other with the actual sifting of those suggestions. In the field of nursing education, suggestions as to possible curricular objectives may be obtained from a number of sources: reports of curriculum committees such as that of the National League of Nursing Education, reports of other groups representing more specialized subject areas, reports of job analysis, follow-up studies of graduate nurses, and studies of student nurses themselves. Reports of various curriculum committees indicate, of course, what these experts consider to be the most worthwhile outcomes of instruction for their particular areas. These reports are valuable because they represent an accumulation of experience as to what is worth teaching. Reports of job analyses are valuable for keeping the curriculum in step with the actual needs of the nurse on the job, and for keeping the curriculum in step with advances in medicine and nursing. Reports of job analyses may also suggest the desirability of setting up fields of specialization within the nursing curriculum to parallel the different kinds of nursing fields. Follow-up studies of graduate nurses can be helpful to the school of nursing in identifying points of strength and points of weakness in professional preparation. Points of weakness are particularly important because they suggest needs that should have been anticipated and met. For example, possible needs might include increased skill in applying principles of mental hygiene in dealing with patients, or knowledge of new techniques arising out of atomic research. In the same way, studies of student nurses may suggest points of strength and points of weakness peculiar to the particular student population, so that the curriculum ought to be adjusted to devote more time to the weaknesses and less to the points at which the students are already quite competent.

Since the consideration of each of these sources will ordinarily result in a list of instructional objectives longer than the individual school of nursing can provide for, it is necessary to

screen these various suggestions to select a smaller number that can be attained by students during the period of their professional education. Here, again, the process should be a rational one; deliberate effort should be made to make choices that reflect a careful consideration of relevant criteria. A first criterion might be the philosophy or concept of nursing held by the particular school. If the school places great value upon health teaching as well as care of the sick, upon mental hygiene as well as physical treatment, then it will try to provide instruction in these areas. On the other hand, if a school adheres to a more conservative definition of nursing it may not feel compelled to provide such instruction. Similarly, the resources of the school act as a screen in determining what kinds of instruction it should and should not attempt to provide. Resources here would include the staff, library, laboratories, equipment, and the like. Another important factor influencing the curriculum is the kind of nurse demanded by the various hospitals which draw upon the school of nursing. It is conceivable that in a certain geographical area there may be much more demand for certain kinds of nurses than for other kinds. This differential demand thus serves as an important factor in determining the kinds of preparation which a school of nursing emphasizes.

The result of this procedure of screening possible objectives will not be uniform. The particular objectives which one school of nursing attempts to attain will differ somewhat from those of another. This variation is not necessarily undesirable; it will reflect, in part, the adaptation of the curriculum to local conditions. Actually, in view of the similarity of nursing activities from place to place, one would expect the various curricula to emphasize many common objectives.

In the process of formulating objectives, it is important that they be listed for each course or comparable unit of instruction. It is not enough to say that we are going to teach human physiology, or mental hygiene, or surgical nursing. A statement of this kind is too vague, for it does not indicate in sufficient detail the particular understandings, skills, and abilities to be developed. It is necessary to go on and actually list the important facts, concepts, and generalizations, and the important skills and abilities that are desired as outcomes of a given course.

Such a listing serves to facilitate the next step, that of defining these outcomes so as to make them measurable. If we are to predict the ability of students to achieve the outcomes of instruction, then we need to have a fairly clear idea of what they are. Take the desired outcome, understanding of and ability to apply principles of mental hygiene. This objective, as it stands, is general and needs to be clarified by listing the specific components which it includes. Some of the important specific behaviors which make up this more general outcome might be the following:

- a Understanding of the basic principles of mental hygiene
- b An objective or scientific point of view toward human behavior—i.e., a recognition that all behavior is caused
- c Ability to distinguish between symptoms and causes of behavior disorders
- d Ability to make a satisfactory tentative diagnosis of common behavior disorders
- e Ability to adapt nursing and therapeutic techniques to patients with differing needs and conditions
- f Skill in helping patients to develop insight into their own problems and to work out their own solutions

When objectives have been analyzed in this manner, they become fairly precise specifications for the development of achievement tests and other means of appraisal, as well as for the detailed planning of the curriculum. Success in the professional school may then be considered as the extent to which students have actually attained these defined objectives at various stages of their professional education. These measures of status or achievement constitute the best criterion measures for validating selection procedures. They are the kinds of performance we are trying to predict.

Ideally, we would like to have some assessment of the student's status and achievement with respect to all of the important objectives at the time of completing the professional program. Since this is likely to be impractical under present conditions, it may be necessary to use as criterion measures achievement at particular stages of the total program. For example, the student's performance at the end of the First Semester may be used as an index of performance in later semesters. First-semester performance is actually a rather good predictor of later performance, particularly when taken in com-

bination with a measure of pre-professional scholarship performance and one of scholastic aptitude. A limitation of first-semester or first year performance as a criterion of success for the entire program is that the type of school work often tends to shift from theoretical to practical courses. This shift in work places somewhat different demands upon the student. However, in determining the usefulness of selection tests, first-semester performance can serve fairly well as a criterion of success for all entrants in the school, including early withdrawal, and as a predictor of success for those students who continue beyond the First Semester. Criterion measures, of course, may be limited to measures of achievement in particular courses.

In addition to these partial criteria of performance, two other criteria should be mentioned, namely, graduation and over-all grade point average. The criterion of graduation vs non-graduation is useful in determining the characteristics which distinguish between students who graduate and students who withdraw or otherwise fail to graduate. The over-all grade-point average may be useful as a single index of success because of its stability and convenience, but its exclusive use in many prediction studies serves only to perpetuate the hackneyed practice of determining the empirical relationship between grades and scores on aptitude tests. In general, these two criteria suffer from the limitation of not being analytic; they do not analyze achievement into its component parts. Broad criteria of this kind are supplementary measures, but not substitutes for separate measures of the degree to which each important goal of instruction has been achieved by students.

Determining the Predictive Measures

The second important problem is that of determining what measures to use in the process of selection. The test of any such measure, of course, is its usefulness in identifying good risks for professional education. A selection or predictive measure is useful to the extent that it indicates a candidate's relative probabilities of success in completing various aspects of the program. This statement implies that the problem is basically

statistical, for its data can be expressed in terms of relative probabilities of success.

Relative probabilities of success are determined empirically by relating two sets of data—(1) criteria of proficiency which become available during the course of professional education or later professional work, and (2) appraisals of the candidate's behavior and characteristics prior to, or early in, professional education. Relationships established empirically on one group are then ordinarily applied to subsequent groups, provided the attendant conditions have not appreciably changed.

Correlation is the usual method for expressing relationships between these two sets of data. Essentially the task is to get predictive measures such that variance on them will be highly associated with variance on the criterion measures. When such correlation is obtained, it is assumed that the two kinds of performance, predictor and criterion, have much in common—the factors that bring about high performance on the predictors are substantially the same factors that bring about success in various aspects of professional education.

A crucial decision, therefore, is that involving the choice of selection tests. Unfortunately, this choice has all too often been made without a careful study of the functions measured by the tests and their relationship to the functions measured by the criteria. Tests have frequently been tried out in a hit-or-miss manner in the hope that they would correlate highly with the various criteria. While all selection tests should be validated empirically through follow-up studies of their effectiveness, it is sound practice to try to obtain validity at the outset through rational analysis of the criteria. A careful analysis of the abilities and traits required for successful performance on the criteria should result in a series of hypotheses as to what the selection tests ought to measure. For example, the curriculum of a medical school will typically include a great deal of course work in the natural sciences. Consequently, one might hypothesize that the ability to read and analyze fairly difficult natural science materials would be an important factor contributing to academic success. In the same way, one could analyze other aspects of the program and attempt to form

rational hypotheses as to the abilities and traits required for successful performance. These hypotheses constitute, in effect, specifications for our selection tests. They provide a rational basis for choosing from among existing tests and for developing new methods of selection.

Another approach to the determination of predictive measures involves a comparison of students who graduated with those who eventually withdrew or failed to graduate. The basic idea is that the graduates possessed certain presumably desirable characteristics at the time of entrance which were not possessed, or were possessed in a limited degree, by the non-graduates. Hence the problem is to identify such characteristics which will differentiate statistically between the two groups of students.

The use of this method does not mean that any and all characteristics found to differentiate should be accepted unequivocally. A characteristic so identified should have some real relationship to success in the professional school and in later professional activity, and should be scrutinized from the standpoint of broader social considerations. For example, the possibility that a higher proportion of graduates of medical schools than of non-graduates had a relative who was a doctor would not necessarily mean that, other factors being constant, preference be given the candidate who was related to a doctor. The statistics themselves may reflect traditional admissions practices. True, relationship to a doctor may frequently be an indirect index of necessary aptitudes and interest, but an important social question remains. This is the question of whether admissions procedures and standards should perpetuate the status quo in the profession by selecting candidates who are quite similar to each other in all of their characteristics and to the active practitioners. The writer will not attempt to answer this question.

Hypotheses as to possible differentiating characteristics may be drawn from such sources as a review of previous investigations, analysis of admissions and other biographical data, and observations of students in counseling and classroom situations. A sample list of hypotheses that might be applicable to the different fields of professional education includes the following

a. The graduate, in comparison to the non-graduate, is more likely to have based his vocational choice upon realistic considerations than upon stereotyped conceptions of the profession or a too lofty idealism. That is to say, he will look upon the field primarily as a worthy occupation for which he is well suited rather than solely as a quick way to get rich, or a glowing opportunity to serve a sick and downtrodden humanity, or some similar consideration. The basic assumption here is that a realistic attitude will enable the student to persist in his education whereas other considerations may be inadequate to overcome basic deficiencies in such factors as health, aptitudes, and a genuine interest in the field.

b. The graduate, in comparison to the non-graduate, is more likely to have had better study habits in preprofessional education. The assumption here is that efficient study habits formed in the earlier years of education tend to persist during the later years and to influence markedly the student's academic success.

c. The graduate, in comparison to the non-graduate, is more likely to have had good health at the time of entrance and during professional education. The assumption here is that good health enables one to work energetically and to withstand the physical strains that accompany work in the professional school and the occupation itself.

d. The graduate, in comparison to the non-graduate, is more likely to have been well adjusted socially and emotionally. His personality was more likely to be stable and free from emotional conflicts. It is assumed here that the well-adjusted person is better able to maintain an even tempo in his studies and withstand the various environmental and emotional forces that disrupt effective learning.

Other hypotheses could be offered in addition to the foregoing. Such characteristics as socio-economic status of family, size of high-school graduating class, extent of participation in extracurricular activities in high school, and marital plans might reveal important differences which could function as indirect indexes of more directly desirable qualities. They should probably be scrutinized from the standpoint of social desirability as well.

It is to be noted that the assessment of these personal and

background characteristics frequently calls for methods of appraisal other than formal paper and-pencil tests. The assessment of health, of course, is best accomplished by a physical examination. Various items of background information such as reasons for vocational choice, father's occupation, and the like, may be obtained from a carefully developed biographical questionnaire. Evaluations of study habits, motivation, and personal-social adjustment may be obtained in the form of appraisals by former instructors, but the alternative methods of obtaining these evaluations are quite numerous.

This approach, in short, provides additional standards useful in the selection process. Many of them are of a non-intellectual sort and therefore complement the more formal measures of intellectual abilities and knowledge. It is in this direction that much future research will profitably move.

Limitations to Prediction

Even though a professional school has established various criteria of proficiency and has utilized valid predictive measures, it will still find that its procedures leave much to be desired in the way of accurate selection. When the correlation technique is used, the term "ceiling" is applied. "Ceiling" refers to the fact that most predictive measures seldom account for more than 50 per cent of the variance in the criterion scores. This generalization holds true for various combinations of aptitude measures as well as for single measures. The addition of one or more other measures to one which already accounts for 40 to 50 per cent of the variance in achievement generally does not raise the correlation appreciably. A ceiling or limit to prediction is reached no matter how many aptitude measures are used or how valid they otherwise appear to be.

Two important implications follow from this fact. The first is that we are somewhat limited in the accuracy with which we can predict academic achievement. While, in general, the students who make the higher scores on the aptitude tests will also make the higher scores on the achievement tests, some of these able students will achieve at only a mediocre level or will eventually even drop out of school entirely and, correspondingly, some of the less capable students will achieve at a level

much above expectations and will complete the program successfully. Obviously, such departures from the expected will limit the usefulness of our selection procedures in identifying good risks.

The second implication is that the remaining 40 to 50 per cent of the variance in achievement not accounted for by measured aptitudes must be attributed to other factors. These other factors are usually considered to be of a non-intellective sort, particularly motivational, but up to the present they have not been accurately defined and measured.

Even if these other important systematic factors were identified and measured, it is quite likely that the accuracy of prediction would be limited. A scrutiny of the learning process is enough to indicate this, for many factors other than aptitudes influence the degree of learning day in and day out: interest in particular subjects, degree of rapport with particular instructors, extent of participation in extra-curricular and out-of-school activities, health, home conditions, financial resources, and marital status, to mention the most important. Furthermore, like many events in life, these factors are to a certain extent unpredictable. And although they may operate quite independently of aptitudes, collectively they are almost as important for they influence the effectiveness with which the student uses his capabilities.

Limitations may inhere in the various criterion measures. This is not surprising in view of the complexity of learning. By and large, the criterion measures now in use—the average grade in a course or program—have a rather limited validity. A grade in a course or program is, after all, a kind of summary evaluation which indicates the over-all success of the student. Such evaluations have some usefulness in prediction studies but, in general, suffer from the limitation of not being analytic, since they do not indicate the extent to which each one of a comprehensive array of desired outcomes has been achieved by individual students. The weakness of an average is that it is always somewhat artificial; it implies uniformity where variability is the rule. Instead of describing the pattern of achievement over the various instructional objectives, it yields only a conglomerate the parts of which are rather non-descript.

Grades and grade point averages commonly suffer from another serious limitation— their meaning varies, they may represent different things in the same course. A given grade may signify one of several things. It may be used to indicate the amount of progress that a student has made in a course, irrespective of the actual level of achievement. Or, it may indicate the amount of effort that the student has put forth, regardless of progress or final status. More commonly, of course, a grade is used as a measure of the final status or level of achievement that the student has reached, regardless of the amount of effort expended or the relative degree of progress made. Now what frequently happens is that a given instructor may be assigning grades to different students on different bases, rather than on a uniform basis. Similarly, different instructors teaching the same course may use quite different bases and standards in assigning grades. Variations of this kind tend to lower the validity of the grades as criterion measures, and consequently reduce the correlation between the predictive and the criterion measures. Unreliability of the criterion measures, of course, also lowers the validity coefficient.

These are not the only problems in connection with the validity and reliability of criterion measures, but they are among the most important. Problems similar to these also arise in attempting to obtain satisfactory predictive measures. They need not be elaborated here.

One other source of difficulty should be discussed. That source is the relative homogeneity or restricted range of talent of the student population. Validity coefficients increase when a test is used on a group with a wide range of aptitude, and decrease when the test is used on a relatively homogeneous, preselected group. Since many students of relatively low aptitude are refused admission to professional schools, the group finally admitted is always more homogeneous in aptitude than the complete group of applicants. Had all of the students who applied been admitted and allowed to continue in school as long as they could, we would obtain higher coefficients of correlation between aptitude tests and school achievement. Practically, of course, there is no point in admitting poor risks if we have good grounds for considering them so.

This rather well-known phenomenon of shrinkage raises an important consideration. Thus, the effectiveness of selection may be judged in terms of the magnitude of the validity coefficients, rather than in terms of the proportion of candidates who have successfully completed the program of the professional school. If we do look at the proportion of successful candidates, rather than at selection procedures only, then certainly we need to consider how effective the total program was in bringing the candidates up to acceptable standards of development. (A great deal happens after a student is admitted to a professional school.) The validity of objectives, the effectiveness of instructional methods and materials, the effectiveness of student personnel services, and the validity of methods of evaluating student progress all need to be reconsidered. This all goes back to the fundamental thesis of this article, namely, that selection procedures should be made an integral part of the total educational program.

Improving Selection Procedures

The thesis of this article is that the planning of admissions or selection procedures should be integrated as closely as possible with the planning of other aspects of the educational program. Those aspects of the program of professional education which are closely related to the planning of admissions procedures are the following:

- 'The selection and definition of instructional objectives
- 'The selection and organization of learning experiences to attain these objectives
- 'The development of measures of the extent to which these objectives have been achieved by students
- 'The development of measures to predict achievement of the objectives
- 'The carrying on of continual research

Operationally defined objectives are the threads which tie together these various aspects of the total program. Objectives are crucial because they function as criteria by which the curriculum is planned, carried out, and evaluated. Stated otherwise, objectives become criteria by which subject matter is outlined, instructional materials selected, instructional methods

developed, and examinations and other instruments of evaluation prepared. A change in the nature of objectives will ordinarily necessitate corresponding changes in the various other aspects of the total program.

The careful formulation and definition of the professional school's objectives make possible the development of examinations and other means of appraisal which will reveal the degree to which these various objectives have been attained by individual students. The scores derived from these tests represent the most directly valid criterion measures that can be obtained.

The analysis of objectives and the development of valid measures of their attainment then make possible a more rational choice of selection procedures. The individual professional school should form hypotheses as to the abilities and traits required of students if they are to complete the program—that is, attain the various objectives. These hypotheses may then serve as tentative guides for choosing aptitude tests and developing other selection procedures. This kind of approach would be a decided improvement over the practice of choosing tests because they happen to be widely used.

Aptitude tests and other selection procedures thus need to be validated against the criteria of proficiency derived from the educational objectives. But selection procedures should not be validated against the criteria solely in terms of validity coefficients, but also in terms of the proportion of candidates who have successfully completed the program. And so the emphasis is as much upon the program of the professional school as it is upon the admissions procedures.

Each professional school should carry on continual research on the effectiveness of its selection procedures and various other aspects of its total program. Selection procedures need to be empirically validated, since one cannot assume that they will be effective in one situation if they have been so in others.

Continual research is also necessary because of the fact of change. Changes may be introduced in the curriculum—new objectives may be added, others revised or dropped; standards of achievement may be raised or they may be lowered. Any one of these will change the criteria of proficiency. Changes

may occur in the nature of the candidates and the student group finally selected. Some professional schools may be attracting many more capable candidates than formerly, perhaps the opposite would be true in other schools. In addition, new developments in techniques of assessing promise and aptitude for professional education may be worth trying out. All of these considerations and many others suggest that the professional school make a periodic check on the effectiveness of its program.

REFERENCES

1. David, Lily M. "The Economic Status of the Nursing Profession " *Monthly Labor Review*, LXV (1947), 20-27.
2. Snygg, Donald, and Combs, Arthur W. *Individual Behavior* New York, Harper Bros., 1949
3. Super, D. E., *Appraising Vocational Fitness* New York: Harper Bros., 1949
4. Thorndike, R. L. *Personnel Selection: Test and Measurement Techniques*. New York: John Wiley, 1949
5. Travers, R. M. W. "Significant Research on the Prediction of Academic Success," *The Measurement of Student Adjustment and Achievement*. Ann Arbor: University of Michigan Press, 1949. Pp 147-190.
6. Tyler, R. W. "Syllabus for Education 360: Basic Principles of Curriculum and Instruction" Chicago: University of Chicago Bookstore, 5820 Kenwood Avenue, 1947. 71 pp (mimeographed)

PREDICTING ACADEMIC ACHIEVEMENT WITH A NEW ATTITUDE-INTEREST QUESTIONNAIRE—I

R. C. MYERS and D. G. SCHULTZ

Educational Testing Service

Early in 1947 the College Entrance Examination Board decided to attempt the development of an instrument for measuring nonintellectual factors associated with subsequent academic achievement in college. Previous attempts to predict academic achievement through the use of attitude, interest, motivation, or personality inventories had uniformly shown a positive but discouragingly small relationship. Ellis (1, 2, 3) has been most active in summarizing results of experiments in this field. Nevertheless, we felt that better success might be obtained by the use of a questionnaire especially designed for inquiring into pertinent attitudes, interests, and motivations of entering college freshmen, previous experimentation having been carried out largely with the use of standard instruments not especially designed for this purpose. Accordingly, we undertook this research for the College Board.¹

Items were constructed covering areas roughly blocked out as: motivation for attending college, intellectual interests, teacher relations and study habits, withholding from outside activities, and parental backing. Altogether, the original questionnaire presented the respondent with 328 questions. This was administered in October, 1947, to the members of the freshman class (Class of 1951) at an eastern women's liberal arts college three weeks after their admission to the college. This first administration showed that, although the question-

¹ The questionnaire was developed, pretested and administered under the general direction of H. S. Conrad, now Chief Research and Statistical Service, U. S. Office of Education. Acknowledgment is due W. H. Schreder and H. L. Green, Jr. of the Educational Testing Service's professional staff for their aid and advice in the analysis of results obtained. The problems involved in using attitude-interest questionnaires for selecting college students were discussed at greater length in a paper describing this study presented by D. G. Schultz during the American Psychological Association Meeting at The Pennsylvania State College, September 3, 1949.

naire was administible, its length made it very unwieldy. Therefore, the sections covering parental backing and withholding from outside activities were arbitrarily dropped in subsequent reproductions. The effect of this action was to reduce the number of similar scorable attitude-interest items between the first and later administrations to 145.

During the remainder of 1947, and throughout the spring and summer of 1948, the revised questionnaire was administered to applicants for the 1948-49 freshman class (Class of 1952) at this same women's college along with their usual application papers.² Questionnaires received from applicants who for any reason were not later actually admitted to the Class of 1952 were discarded.

Allowing the Class of 1951 to complete their freshman year, the initial task we set for ourselves may be summarized as follows:

1. Give each member of the Class of 1951 a relative achievement rating or index, i.e., an index of freshman-grade performance relative to scholastic aptitude.

2. Compare the questionnaire responses of overachievers with those of underachievers in order to find those items showing the greatest response differences.

3. Compare the questionnaire responses of the Class of 1951 with those of the Class of 1952 in order to determine what effect the difference between postadmission administration (Class of 1951) and preadmission administration (Class of 1952) had on average responses.

4. Prepare a key by which to score the Class of 1952 questionnaires, based upon overachiever-underachiever response differences found in the Class of 1951, and adjusted for preadmission-postadmission response differences that may have been found to affect keyable items.

5. Using this key, score the Class of 1952 questionnaires, thus obtaining a predictor of academic achievement for the members of this class.

The "Achievement Index."—All applicants for admission to the

² The questionnaire has also been administered under similar conditions to applicants at this college for the 1949-50 freshman class (Class of 1953). However, in this paper we are interested only in the first two administrations.

college where we carried out our study are required to take the *Scholastic Aptitude Test* of the College Entrance Examination Board. These scores on the SAT verbal and mathematical sections compared to Freshman-Year Grade-Point Averages of the members of the Class of 1951 formed the basis for assigning an achievement index to each student. The following regression equation was used in calculating the most probable Freshman-Year GPA from the SAT-V and SAT-M scores, $\hat{Y} = .002(X_1 + X_2) + 4.695$, where $X_1 = \text{SAT-V}$ and $X_2 = \text{SAT-M}$. The multiple correlation of SAT-V and SAT-M with Freshman-Year GPA was found in this case to be .51. The intercorrelations of these measures are shown in Table 1.

TABLE 1
Intercorrelations of CEEB Scholastic Aptitude Test Sections and Freshman-Year Grade-Point Average, Class of 1951 (N = 355)

Measures	SAT-V	SAT-M	Freshman-Year GPA*	Mean	S.D.
SAT Verbal		.70	.44	505	.93
SAT Mathematical	.77		.39	480	.75
Freshman-Year Grade-Point Average*	.44	.39		2.53	0.57

* Throughout this article, the signs of correlations with high-school averages and with college averages have been changed so that a positive correlation here indicates that high achievement was associated with high values on the other variable.

Each student was then assigned an achievement index on a standard scale having a mean of 13 and a standard deviation of 4. Those whose test scores and GPA brought them furthest below the regression line received the lowest achievement indices and were by definition underachievers, while those furthest above the line received the highest achievement indices and were by definition overachievers.

Overachiever-Underachiever Differences —In order to accentuate whatever differences might exist in the characteristics and questionnaire responses of overachievers and underachievers, it was decided to compare the 37 most extreme overachievers with the 37 most extreme underachievers. It is interesting to note from the figures presented in Table 2 that our overachievers had also been overachievers in high school. Although their scholastic aptitude scores are somewhat lower than those

of the underachievers, their high-school grades were significantly higher. (Since Freshman-Year College GPA is undoubtedly very highly correlated with the achievement index, the large *t*-value for the difference between the groups on this variable is really meaningful only in demonstrating that the extreme groups selected were actually quite disparate.)

TABLE 2
Comparison of Extreme Overachievers and Extreme Underachievers on SAT Scores and Academic Achievement, Class of 1951

Measure	Extreme Overachievers (N = 37)		Extreme Underachievers (N = 37)		Difference	<i>t</i>
	Mean	S D	Mean	S D		
SAT-Verbal	505	101	520	94	15	0.65
SAT-Mathematical	471	90	487	60	16	0.89
High School Average, Adjusted*	1.32†	0.28	1.84†	0.45	0.52	5.89
Freshman Year College Grade Point Average	1.75†	0.35	3.39†	0.39	1.64	18.92

* Adjusted by the college on the basis of previous experience with graduates from those high schools concerned

† On this variable the higher the numerical value the poorer the achievement represented

TABLE 3
Age at High School Graduation of Extreme Overachievers and Extreme Underachievers Class of 1951

Age (Years and Months)	Number of Extreme Overachievers	Number of Extreme Underachievers
18-6 through 18-11	2	1
18-0 through 18-5	6	9
17-6 through 17-11	12	14
17-0 through 17-5	8	10
16-6 through 16-11	6	3
16-0 through 16-5	3	0
Total	37	37

The data on age at high-school graduation and size of high-school senior class shown in Tables 3 and 4 are presented for their suggestive usefulness only. We are not prepared to state categorically that overachievers will generally be found to have graduated from larger high schools and at a younger age than underachievers. Nevertheless, such a conclusion does not seem unreasonable when such factors as grade-skipping and greater competitiveness in larger classes are considered.

Each of the 145 questionnaire items had been provided with multiple choice responses ranging from negative to positive. These possible responses could be considered as lying on a continuum whose length was determined by the number of equal-appearing interval responses that had been provided. As a first step in obtaining items to be utilized for a scoring key, it was

TABLE 4
Item 1 High Achiever, Item 2 C 22 of Extreme Overachievers and Extreme Underachievers, Class of 1951

High and Low Achiever Class 22	Number of Extreme Overachievers	Number of Extreme Underachievers
200 and over	19	15
100-199	16	10
50-99	1	7
25-49	0	4
Under 25	1	2
Total	37	37

TABLE 5
*Hypothesis 1 of Item Associated with Unpaired Chi Squares for 62 Selected Items 37
 Extreme Overachievers to 62 Extreme Underachievers, Class of 1951*

P	Number of Items	Cumulative Number of Items
< .1	2	2
< .1 > .2	1	3
< .2 > .3	1	6
< .3 > .4	4	10
< .4 > .5	9	19 ¹
> .5 < .6	12	31 ¹
> .6 < .7	14	45 ¹
> .7 < .8	10	55
> .8 < .9	6	61
> .9 < 1.0	1	62

¹ used for first scoring key

² used for second scoring key

³ used for third scoring key

decided to select those in which the average response difference between overachievers and underachievers amounted to at least 5 per cent of the total continuum length. This procedure resulted in the selection of 62 items. The overachiever and underachiever responses to these 62 items were then compared by the Chi-square technique, resulting in the distribution of P values shown in Table 5.

Discarding those items having a P of $> .50$ left us with a

maximum of 45 items with which to work out a scoring key. Actually, three scoring keys were eventually used in scoring the Class of 1952 questionnaires: one based on the 45 items with a P of < 50 associated with Class of 1951 overachiever-underachiever response differences, a second based on the 31 items with a P of < 30 , and a third using only the 19 items with a P of < 20 . This method which we adopted for selecting the questionnaire items is essentially similar to a method which Thorndike has described (6).

Preparatory to constructing the scoring keys, however, it became necessary to make allowance for the important difference in the social situations under which the members of the Class of 1951 and of the Class of 1952 had answered the questionnaire, the first group responded after they had achieved their goal of college admission, while the second group responded before they were admitted and even before they knew whether or not they would be admitted. As was expected, this difference in the conditions under which the questionnaire was administered to the two groups was evidenced by a great deal of disparity in average response of the groups to the same questionnaire items. Of the 145 items, 90 were found to have significant preadmission-postadmission response differences (CR of 2.00 or higher). This finding in itself has been considered of sufficient interest to be reported upon separately (4). For present purposes we will simply give an example of how the scoring key was modified or adjusted to take account of the difference in average preadmission and postadmission response.³

Item No. 113 asked: *Do you feel that the higher the grades a girl gets in college the more she will amount to after college?* The five responses that were provided for this item were (1) *Yes*, (2) *Probably Yes*, (3) *Uncertain*, (4) *Probably No*, (5) *No*. The

³ The unadjusted scoring key for the 45 items with a P of < 50 was applied to the questionnaires of the 74 extreme achievers in the Class of 1951 on whom the key was based. In order to determine the relationship between this score and the achievement index, a biserial correlation from widespread classes was computed, the standard deviation of the questionnaire scores for the complete population being inferred from the two tails by means of a formula given by Peters and VanVoorhis (5). Inclusion of only the widespread classes made unnecessary the scoring of the questionnaires of all 355 students in the Class of 1951, a task which seemed unwarranted in view of the fact that the key was being used with the same group on whom it had been based. This correlation coefficient between questionnaire score and achievement index, an estimate of the value for the full population of 355, was $+ .36$.

average preadmission response (Class of 1951) to this item by the overachievers was 3.40, and by the underachievers was 3.73. For the entire Class of 1951 the average response was 3.62, but for the preadmission group (Class of 1952) the average response was 2.72. Thus, in preparing the scoring key, a shift of the preadmission group toward the "Yes" terminus of the response continuum had to be taken into account. This shift, or average difference, in this case amounted to .90 or 22.5 per cent of the entire length of the continuum (90/4.00). Presented below is the scoring key that would have been used if the preadmission-postadmission difference had not been considered, as compared to the adjusted key which was used:

Score to be given for indicated response	0	2	3	5
Unadjusted Key	5	4	3	2 or 1
Adjusted Key	5 or 4	3	2	1

As is evident from the above, the key was so devised that the higher the score the greater the prediction of overachievement; also, so as to give additional weight to extreme responses. Using the key, each of the 346 Class of 1952 questionnaires was given three different scores, as previously explained, on the basis of the Class of 1951 overachiever-underachiever P values of the differences. The means, sigmas, and reliabilities of these three scores are shown in Table 6. The rather low split-half reliability coefficients obtained indicate that the items included are fairly heterogeneous in nature.

In July, 1949, after the members of the Class of 1952 had completed their first year, we received their Freshman Grade-Point Averages and were able to proceed with the preparation of the criterion, the achievement index. This was prepared in precisely the same manner as had been the case with the Class of 1951, namely, by regressing Freshman-Year GPA on CEEB Verbal and Mathematical Aptitude Test Scores. The formula used in this instance was $\bar{Y} = .002(X_1 + X_2) + 4.610$, where $X_1 = \text{SAT-V}$ and $X_2 = \text{SAT-M}$. The multiple correlation of SAT-V and SAT-M with Freshman-Year GPA was .52. The intercorrelations of these measures are shown in Table 7, which may be compared to Table 1 for an indication of the stability of

the measures at this college for the two successive freshman classes

Table 8 shows that, using the key on the Class of 1952 questionnaires, the correlations obtained for the three scores with the achievement index were .12, .10, and .14, respectively. Thus, in our first try-out of this new questionnaire we had predicted by this extent the first year academic achievement (grade performance relative to scholastic aptitude) of our experimental group

It will be noted in Table 8 that all of the three scores had correlations of essentially zero with adjusted high-school aver-

TABLE 6
Questionnaire Scores, Class of 1952 (N = 346)

	Score 1 (45 items)	Score 2 (31 items)	Score 3 (19 items)
Mean	128.1	88.1	57.6
Standard Deviation	14.9	10.0	8.5
Reliability (odd-even correlation, corrected)	.52	.41	.45

TABLE 7
Intercorrelations of CEEB Scholastic Aptitude Test Sections and Freshman-Year Grade-Point Average, Class of 1952 (N = 346)

	SAT V	SAT M	Freshman Year GPA	Mean	S D
SAT-Verbal		.36	.47	515	92
SAT-Mathematical	.36		.39	494	80
Freshman-Year Grade Point Average	.47	.39		2.51	0.57

age and with the mathematical aptitude test; scores 1 and 3 also show zero correlations with the verbal aptitude test. This is definitely encouraging. Whatever aptitudes are being measured by the questionnaire we can be reasonably sure are other than those currently measured by the more traditional measures.

The figures in Table 9 indicate the effect of adding Questionnaire Score 1 to the other variables in predicting Freshman-Year College Grade-Point Average for the Class of 1952. Including the questionnaire score with the two SAT scores increased the multiple correlation from .52 to .53, while adding it to the battery of SAT scores and adjusted high-school average increased

the multiple correlation from .63 to .64. It is of interest to note that addition of the questionnaire score did not essentially change the beta weights of the other variables and also that the beta weight of the questionnaire score was almost the same whether or not the adjusted high school average was present in the predictor group. These facts tend to confirm the point

TABLE 1

Intercorrelations of Predictor Variables with Freshman Achievement and SAT Scores, Class of 1952, $N = 280$

	Multiple Correlation		
	1 (SAT)	2 (SAT + Average)	3 (SAT + Average + Questionnaire)
Adjusted Freshman Achievement	.62	.64	.64
Freshman Year College Course Point Average	.0	.11	.11
SAT Verbal	.2	.18	.02
SAT Mathematical	.12	.1	.03
High School Average Adjusted	.2	.47	.22
Questionnaire			
Score adjusted		.2	.82
Score adjusted	.3		.93
Score, adjusted	.64	.71	
Mean	572.6	574.1	576.6
S.D.	118.9	111.1	115

TABLE 2

Multiple Correlations of Various Combinations of Predictor Variables with Freshman Year College Course Point Average, Class of 1952, $N = 280$

1	Predictor Variables		2	Multiple Correlation of Freshman Year College Course Point Average with Predictor Variables	
	SAT	High School Average		SAT	High School Average
1	.63	.232	.634		.52
2	.674	.247			.51
3	.676	.244	.64	.674	.57
4	.674	.252		.64	.54

that the questionnaire score is a member of the battery which is independent of the other predictor variables.

We feel that the difference in time of administration may have made the key based on the 1951 Class (answering after admission) less appropriate for the responses of the 1952 Class (answering before admission) than might have been true if both groups had responded under similar circumstances. The rather low correlations obtained in the first experiment may

have been partially a result of this fact. We are now in the process of preparing a scoring key for the Class of 1953 questionnaires (entered college fall of 1949) from this same women's college. The key will be based upon overachiever underachiever questionnaire response differences in the Class of 1952. No rough and-ready adjustments will be necessary in this case to attempt to compensate for differences in time of administration, as both classes (1952 and 1953) answered the questionnaire before admission and hence were in the same social situation and under the same pressures. The results of this second experiment will be available after the end of the 1949-50 academic year and will be reported upon at that time.

REFERENCES

1. Ellis, Albert. "The Validity of Personality Questionnaires." *Psychological Bulletin*, XLII (1946), 424-440.
2. Ellis, Albert. "Personality Questionnaires." *Review of Educational Research*, XVII (1947), 47-62.
3. Ellis, Albert. "Intelligence and Achievement." *Review of Educational Research*, XVII (1947), 7-24.
4. Myers, R. C. "A Study of Readiness." *Journal of Educational Psychology*, XLII (1949), 149-156.
5. Peters, C. C. and Van Aken, W. R. *Personality Procedures and Their Mathematical Basis*. New York: McGraw-Hill Book Co., 1948. pp. 484.
6. Thorndike, R. L. *Personality*. New York: John Wiley & Sons, 1942. pp. 24-25.

PATTERNS OF RESPONSE IN LEVEL OF ASPIRATION TASKS

LOUIS D. COHEN

Duke University

FERDINAND Hoppe, one of the earliest workers with level of aspiration, in discussing changes in goals following success and failure, indicated that the direction and type of goal adjustment are influenced by the true inner aims and aspirations of the subject (11). Frank, seeking quantification of the concept, defined it operationally as "the level of future performance in a familiar task which an individual, knowing his level of past performance in that task, explicitly undertakes to reach" (7, p. 119). However, as Rotter points out (17, p. 467), Frank's explicit factors were not necessarily the same as the implicit ones in which Hoppe was interested. According to Gardner (8, p. 62) his view may rather be considered the empirical coordinating definition of Hoppe's concept.

As applied in most experimental situations the level of aspiration is determined by asking the subject to perform some task ranging from a simple motor one, such as throwing quoits, to a complex one, such as solving arithmetic problems. Scores on such tasks are made known to the subject and he is asked to state what he "expects" to make next time. The difference between this statement and the preceding achievement has been described as the goal-discrepancy score, and when computed for a number of trials, and the mean taken, such scores are called the average goal discrepancy.

This score has been the most commonly reported aspect of behavior in level-of-aspiration tasks. But it has also become evident that the test situation in which level of aspiration is determined lends itself to other observations. Not alone the average goal-discrepancy score, but the consistency with which goals are set at a high or low level, the frequency of shift in goal level, the sequence of changes in goal level to success and

failure, difficulties in setting goals, the direction of shift in goal level after success and failure—all have been reported as additional aspects of level of aspiration behavior (14).

Level of aspiration has thus tended to be used as a concept describing a complex of behavior involved in a goal setting situation.

Recognition of the level of aspiration as a test situation from which a number of observations could be made led to the formal description of behavior with the task in terms of "patterns" of response. The first major contribution along the lines of pattern was that of Sears, who used the average goal-discrepancy score to characterize her groups (21, 22). She described four patterns: (1) low positive discrepancy score, (2) high positive discrepancy score, (3) negative discrepancy score, and (4) a mixed group in which no clear pattern of reaction was maintained. In elaborating the behavioral correlates to these patterns she observed that the low positive discrepancy-score group included only those who were markedly secure in their achievement, that the high positive discrepancy-score group included those who were able to acknowledge rather freely their own relative incompetence along certain lines, and the negative discrepancy score group included those who found it difficult to admit to another person that they were striving for more than they were able to achieve.

Sears used more than goal discrepancy score in characterizing her subjects. We may note her observation that, "it seems reasonable to suppose that the aspiration level response forms part of a cluster of associated personality attributes which may function as a whole in a number of different situations" (22, p. 335).

Rotter, by extending the description of patterns, brought into clearer focus the need for an evaluation of the variables in the level of aspiration situation (20). He described nine patterns of response: (1) low positive D score¹ pattern, (2) low negative or very slightly positive D score pattern, (3) medium high D score pattern, (4) achievement follower pattern, (5) the step pattern, (6) very high positive D score pattern, (7) high nega-

¹ D score is used by Rotter to indicate average goal discrepancy score.

tive D score pattern (8) rigid pattern, (9) the confused or breakdown pattern

As can be noted by inspection of Rotter's patterns, he has added a characteristic to Sears' goal discrepancy score by including patterns describing the manner of shifting or reacting to success and failure in the level of aspiration task. Patterns 4 (achievement follower), 5 (step), 8 (rigid), and 9 (confused or breakdown) are concerned with describing the method of adjusting goals to success and failure, and say nothing about the height of goal level setting. For example, the rigid pattern (pattern 8), which was characterized by Rotter as involving few if any shifts in goal level, could have set goals either high or low. In the first case the patterns would be rigid *and* low negative D score (patterns 8 and 2), and in the other case they would be rigid and medium or very high positive D score (patterns 8 and 1 or 6). Presumably the behavioral implications would be different in each case.

Duffy, in attempting to establish a systematic framework for the description of personality, has suggested that "personality be described in terms of the direction and the energy mobilization of response, since all behavior shows variation in goal direction and in intensity . . ." (6, p. 189). She goes on to observe that "the broad outlines of personality can be usefully sketched in terms of what the individual approaches or withdraws from, with what intensity, and with what consistency" (p. 189). These observations are appropriate to our analysis of the variables in the level of aspiration tasks.

In the light of Rotter's attempts at pattern formulation and Duffy's concept of two major variables of behavior, it seemed probable that level of aspiration behavior would be more fully described by the use of two major variables rather than one.

The problem set for this paper was the examination of behavior in a level of aspiration task with a view to examining the specific measures used, and defining the patterns of response that might emerge from an integration of the two major variables suggested by the literature: (1) goal level setting (height of average discrepancy score), and (2) method of adjusting goal levels to success and failure.

Subjects, Task, and Scores

Subjects — We selected as subjects for this study a number of patients from the outpatient and inpatient clinics and wards of Duke Hospital. The patients were primarily those who had been referred to the neuropsychiatric department for consultation or treatment, but also a large number of medical patients referred for psychosomatic study to the psychosomatic clinic of the department of medicine and neuropsychiatry.

Group Criteria. The criteria for selection were as follows: white patients between the ages of 16 and 55. Additional con-

TABLE 1
The Experimental Group, Showing Distribution of Diagnoses, Sex, Mean Age for Sex
N = 50

Diagnosis	Male	Female	Total
Psychoneurosis	8	2	10
Schizophrenia	6	0	6
Manic depressive	1	0	1
Anorexia nervosa	0	1	1
Hypertension	4	15	19
Asthma	4	5	9
Ulcerative colitis	0	1	1
Diabetes mellitus	0	1	1
Abdominal pain	0	1	1
Pregnancy, delivered	0	1	1
Totals	23	27	50
Mean age	29.56	34.59	32.28

siderations were: dividing the group into equal numbers of male and female; providing for approximately the same age distribution in the male and female groups, providing for a distribution of diseases of approximately the same type for the male and female groups.

Selection of Patients — Study of Table 1 reveals that in a number of aspects we failed to attain the criteria listed above. Our male group is largely the type of patient found in psychiatric clinics, the female largely the type found in medical clinics. The male group also included a large number of veterans and, in general, was somewhat younger than the female group.

The group as a whole, nevertheless, seems to be representative of those diseases reported as having psychological attributes either etiologically or symptomatically (3). This would

seem to have been an advantage, since our aim in the study of this group was to get at certain modes of adjustment which might have been more dramatically in evidence in a patient group with emotional difficulties than in another group in which such differences were less obvious.

Task. Among the various level of aspiration techniques described in the literature (10, 24), the target-aspiration board used by Rotter seemed particularly appropriate as a clinical instrument. The board is easily carried, can be set up near a patient's bed if he is nonambulatory, does not require much energy output, is interesting and challenging, and seems to evoke active participation on the part of the patient. In addition, considerable opportunity for quantification is possible because of the wide variety of measures that have been suggested and used.

The board we used followed Rotter's specifications. This is a board some 38" long with a square groove down the center. A steel ball is put along the groove by a stick resembling a miniature billiard cue. Regularly spaced depressions preceding the numbered units, and also one placed in the center of each numbered unit, slow down the speed of the ball and provide a resting place for it when it comes to a stop. The score is dependent upon how close to the central unit the ball comes to rest, regardless of the direction. The central unit, painted in white with the black number ten on it, counts 10 points. The ones next above and below count 9 points, the next ones 8, and so on down to 1, the value decreasing as the distance from the 10 increases. The units under 10 are painted alternately blue and gray (18, p. 413, 414).

Rotter calls for the use of a one inch steel ball, but because of the postwar difficulty in getting supplies we had to be content with the use of a $\frac{1}{2}$ " ball. This had the effect of making control of the ball more difficult and consequently of making the problem for the subjects more difficult. It was felt that this additional stress was advantageous to our experiment and did not contravene Rotter's criterion that the problem be within the level of achievement and not so difficult as to make the subject feel "at one extreme of talent without having any immediate comparison with others" (18, p. 411). In analyzing the results on the level-of-aspiration task for the fifty subjects, it was found that there were 497 successful trials and 503 unsuccessful trials. This would suggest that the subjects did not find the task either unusually difficult or easy.

The instructions used were identical with those described by Rotter: After appropriate trials, the subject was asked to

state the score (*aspiration bid*) he expected to achieve in the next series of five trials. The results attained (*performance*) were penalized by giving no credit for achievement above the aspiration and deducting two points for every point that the achievement fell below the aspiration. This score was known as the *earned score*, which was conspicuously announced to the subject. The task continued for twenty series with a rest period after the tenth. After the twentieth series the subject was interviewed as to his reactions to the test.

Scores.—The scoring was accomplished through the use of certain quantitative relationships suggested by Rotter and Klugman (12, 13) and some that were original to this study. However, subsequent analysis suggested that different measures were apparently measuring the same thing. The measures we began with were as follows:

Ds—the mean difference between the earned score and the immediately subsequent aspiration bid for the series of twenty trials (The actual N here is 19.)

Dp—the mean difference between the actual performance and the subsequent aspiration bid for the series of twenty trials

Ds1—the mean difference between the earned score and the subsequent aspiration bid for the first ten of the twenty trials

Ds2—the same as Ds1 for the second ten of the twenty trials

Dp1—the mean difference between the actual performance and the subsequent aspiration bid for the first ten of the twenty trials

Dp2—the same as Dp1 for the second ten of the twenty trials.

successes—the number of times in twenty that the subject's performance equalled or exceeded the aspiration bid.

failures—the number of times in twenty that the subject's performance was below the aspiration bid.

Direction of shift after failure or success—the number of times a subject raised, lowered, or used the same aspiration bid after success or failure

shifts—the number of times in twenty that the subject changed his aspiration bid (e.g., 20, 25, 20, 20, 25 = 3).

unusual shifts—the number of times the subject lowered his aspiration bid after success or raised it after failure. Note was also made of the different situations, i.e., after success or after failure.

repressions—the number of times the subject failed to announce his aspiration bid before beginning a new series.

J score—judgment score—the difference between the aspiration bid and the actual performance. (Attainment direction. (Attainment

- C_1 shift — the number of shifts converted to percentages
 $(N = 19)$
 C_2 confirming shift — number of shifts in which the subject raised his aspiration level after success and lowered it after failure, converted into percentage
 C_3 same as C_2 absence of shift — number of times in which subject failed to change his aspiration level, converted into percentage
 C_4 deviant shift — number of shifts in which the subject lowered his aspiration level after success and raised it after failure, converted into percentage
 $\#$ different levels — the number of different numbers used as aspiration levels (e.g. 20, 23, 25, 27, 29 = 5)

In addition to these measures certain other relationships between measures were investigated:

- D_1 D_p — the difference between D_1 and D_p
 D_{1-1} D_{1-2} — the difference between the first and second halves of D_1
 D_{p1} D_{p2} — the difference between the first and second halves of D_p
 D_{1-1} D_{p1} — the difference between the first half of D_1 and D_p
 D_{1-2} D_{p2} — the difference between the second half of D_1 and D_p

Analysis of Scores

The D_1 score — The traditional scores used in level of aspiration studies are our ' D_p ' (goal discrepancy score) and the "number of shifts" score. However, in examining the test situation we noted that the subject, after the completion of a series of trials, was confronted with two scores, the first, his actual performance score, and the second, the earned score. While the earned score was emphasized by the examiner there was no assurance that it was of equal importance to the subject. We felt it highly probable that the subject might react to the performance score rather than to the earned score, or perhaps to some combination of the two. We therefore used both a D_1 and D_p score for each subject. However, in computing the Pearson product moment r for these two variables we secured an r of .948 for our group. It thus seemed that the two measures served identical purposes and we dropped the use of the D_1 score.

D_{1-1} D_{1-2} D_{p1} D_{p2} — The differences between the first and second halves of the test as far as D_1 and D_p scores were con-

cerned were not remarkable. However, there is apparently greater variability in the second half, as noted by the larger SD scores in Ds2 and Dp2. This is also suggested by correlations of .611 between the Dp1 and Dp2 scores, and of .649 between the Ds1 and Ds2 scores.

An examination of scatter diagrams suggests that only a small number of subjects varied between the first and second halves, but that these cases varied markedly. Apparently something caused this marked change in behavior from one part of the test to the next for these subjects.

The differences between the means are not significant ($t = .101, .156$), but the differences between the SD's show significance at the 1% level of confidence for the Ds score ($t = 2.67$) and at the 10 per cent level of confidence for the Dp score ($t = 1.86$).

These findings suggest that there is greater variability in the goal-setting activities in the second half of the test than in the first half. This differs from Rotter's results, which show greater variation in the first half, presumably as a function of the amount of time it takes the subject to find his level. Where Rotter obtained a decrease we found an increase in variability. One possible explanation may be the tendency of some subjects to tire of the task about halfway through, or to become too tense. Some expressed a wish to terminate the session at about the fifteenth series. It may be that at that point the characteristic behavior gave way to other influences and the subject utilized another type of behavior. Other possibilities may have to do with differences in our population or task from those of Rotter.

Successes - As Rotter suggested, and contrary to Klugman, we found a very high correlation between the goal-discrepancy score (Dp) and the number of successes and failures. Since the number of trials is fixed, successes and failures are reciprocal numbers. Our correlation between the number of successes and the Dp score was $-.934$, which agrees with Rotter's findings of .90 and .88.

Thus our Ds, Dp, number-of-successes, and number-of-failures scores give us essentially the same information. This suggests that the relative height of the aspiration settings has the

function of controlling the successes. The higher the aspiration score the lower the number of successes, the lower the aspiration score the higher the number of successes. This very relationship recapitulates the previous findings in level of aspiration studies: (1) low goal setting for the purpose of attaining success, and (2) high goal setting in which possibly the *statement* of goals is itself the goal.

* *Shifts*—Turning now to the number of shifts, we find a very low correlation ($r = .14$) with Dp and a correlation of .214 with the number of successes. It thus appears that there is very little *linear* relationship between Dp, success, and the number of shifts. However, we examined the relationship between the number of successes and the number of shifts for curvilinearity and found an r of .511. Testing the significance through χ^2 , we found a χ^2 of 13.185 with a P of .01, suggesting the probability that the curvilinear relationship held at a confidence level in excess of 1 per cent. Using the χ^2 table of Peters and VanVoorhis (16) we obtained an α of .1921, which also exceeds the one per cent confidence level. The curvilinear relationship noted here seems to warrant special attention.

Our results suggest that both those subjects who shift a great deal and those who hardly shift at all have the greatest number of successes, that those who shift the mean number of times range from the mean number of successes to slightly below. Thus the number of shifts tends to distinguish three types of behavior: (1) Those who shift hardly at all. This is a somewhat rigid type of behavior and may have two results as regards the number of successes. If goals are set low and maintained there, the number of successes is high, if goals are set high the number of successes is low. Our group contained mostly the low goal setters. (2) Those who shift a great deal. This type of behavior conforms to every change in outer pressure. The attempt to attain success by changing goals reflecting the immediate status of achievement tends to insure a high number of successes. (3) Those who shift within the range of ± 1 SD of the mean number of shifts. This behavior apparently involves holding a self level of competence as a goal and reacting to this level, shifting up or down depending upon the *trend* of the performance rather than on the momentary success or failure.

Direction of Shift We were particularly interested in the subjects' reactions to success and failure as they influenced goal setting, and considered three possible reactions: raising, lowering, or holding to the same aspiration level. We felt that by taking the percentages of those reactions that conformed to the previous performance (conf %—raising after success and lowering after failure) a useful measure might be secured (The realism score of Adams [1]). We also designated two similar measures: the percentage of deviant shifts (dev %—raising after failure and lowering after success), and the percentage of absence of shift (same %) which we tentatively related to rigidity of behavior. The combined conforming and deviant

TABLE 2
Correlations Between Certain Measures of Level of Aspiration Performance in the Experimental Group N = 50

	N Sucs	N Repr	Conf %	Same %	Dev %	% Shifts	J
Dp	4.1	1.7	1.4	2.1	1.0	1.4	.217
% Succ		1.4	1.6	2.4	2.2	1.4	-.160
% Repr			1.1	1.9	1.6	1.9	.113
Conf %				.918	.113	.880	-.056
Same %					-.587	-.999	-.058
Dev %						.587	-.021
% Shifts							-.058

$r = .44$ $t = 1.95$ $df = 49$
 (For table of significances see 9, p. 299)

percentages were, of course, equal to the total percentage of shift.

As noted in Table 2, the correlation of these measures with Dp and % successes is not significant (9, p. 299). While the correlations of conforming %, same %, and deviating % with the number of shifts are high, these relationships conform quite closely to expectations. The conforming percentage will be highly correlated with the percentage of shift, since it accounts for all shifting except the deviant. Our correlation of .880 confirms this. The correlation of deviating % with % shift is .587, and seems unusually high. This suggests that those who shift excessively have a greater tendency to shift inappropriately. Our three measures (% conforming, % deviating, % same) do not seem related to our Dp and % successes measures, but rather to the number of shifts.

* *Repressions*—Another measure of interest is the number of repressions. This reflects the anxiety the subject may have regarding goal setting (20, 23). No significant relationships with the other measures were found, with the exception of a relationship to the conforming $\%$ ($r = .33$, significant at the 5% level of confidence). The relationship we found is therefore considered noteworthy. The implication seems to be: the higher the number of appropriate shifts, the lower the amount of concern over choice of goal setting.

J score—The judgment score (the actual deviation in performance from aspiration level) seemed to bear no particular relationship to any of the measures heretofore reported. It may be that the J score is measuring something quite different.

Recapitulation—We ended with the following measures which had some apparent value for our study: Dp, # successes, # repressions, conforming $\%$, same $\%$, deviating $\%$, % shifts, and the judgment score. There seemed to be a grouping of scores measuring the same thing. The Dp and # successes seemed to relate to goal level setting; the % conforming, same $\%$, % deviating, and % shifts seemed to relate to the method of adjusting goals to success and failure; the # repressions seemed also to be related to the method of adjusting goals to success and failure, but to a limited degree, and the J score seemed to be an independent value.

Patterns

Rutter, and Sears before him, pointed out, however, that some of the unique characteristics of the performance in the level of aspiration situation were lost by taking the means of the results of the various measures, since wide individual variation in the test situation was frequently lost in this process. In order to offset the extreme variations that would obscure trends in the performance, rating scales were developed to extract the meaning of the performance from the data.

Rating Scales Developed for this Study—We noted that goal-level setting behavior varied from high positive goal discrepancy scores (Dp) to high negative goal discrepancy scores. In order to distinguish behavior groups this continuum was divided into seven parts on the basis of standard deviations of the means of

Dp and # successes. In considering *methods of adjusting goals to success and failure* we characterized five types of behavior (rigid, arbitrary, flexible, conforming, and achievement following), four of which seemed evident from our data, and one, achievement follower, borrowed from Rotter, which we felt might occur in some subjects. Thus, each subject's performance could be rated for goal level setting and method of adjusting goals to success and failure.

Goal Level Setting. The rating scale for goal level setting was established by dividing the various scores for height of goal level into seven step intervals, based on the standard deviations of the experimental group's performance. Each step interval was described in terms of the behavior that seemed relevant for it, in order to make the rating more than a clerical checkoff, and to make it possible to evaluate properly the few extreme scores that could seriously distort the mean of any single record. The seven step intervals had these descriptions: (1) very high positive, (2) high positive, (3) positive; (4) plus minus, (5) negative, (6) high negative, (7) very high negative. A sample of the ratings, that for the high positive step interval, read as follows:

2. Setting goals considerably above achievement level, goal setting perhaps becoming the goal itself, considerable expenditure of effort to do well, earnest desire to improve performance almost without reference to the actual performance.

Other steps of goal level setting were similarly described.

Method of Adjusting Goals to Success and Failure. - From our observations it appeared that this concept involved a number of variables which we defined operationally. The variables we used were: (1) responsiveness, (2) conformity to the stated goal objectives, and (3) appropriateness of action.

By responsiveness we meant the reaction of the subject to the situation he was in and to changes in his situation. High responsiveness meant that each change in the situation was reacted to by attempts to deal with or to handle it in some way. Low responsiveness meant that changes in the situation were ignored or were dealt with by avoidance.

By conformity to the stated goal objectives we meant responding to instructions by carrying out their intent. High con-

strategies involved adhering to the rules and the framework of the task. Low appropriateness involved changing the rules and framework of the task.

High appropriateness of action involved responding to the situation by reference to previous success and failure in the task. High appropriateness involved responding by action in keeping with previous success and failure in the task, low appropriateness involved responding by action devoid of basis on previous success or failure in the task. The subject responding by inappropriate behavior.

These variables were interrelated and the following charts prepared, showing the values of the variables for the five methods of adjustment we had denoted* (Tables 3, 4).

TABLE 3
Interrelationships Between Factors Used in Rating Methods of Adjusting Goals to Success and Failure

	responsiveness to task	conformity to stated objectives	appropriateness of action
Rigid	low	low	low
Arbitrary	high	low	low
Flexible	medium	high	medium
Adherence to planning	high	low	high
Goal setting	high	high	high

TABLE 4
Typical Scores on Index of Response to Test Situation for Rating Methods of Adjusting Goals to Success and Failure

	flexible	nonconforming %	success %	deviant %
Rigid	2.7	6.5	43.1	0.11
Arbitrary	1.1	6.1	11	37
Flexible	4.0	17	42	5
Adherence to planning	1.9	2.4	5	0
Goal setting	0.1	7.1	26	0

The description of the ratings follows:

1. **Rigid** - Low responsiveness to test situation, low number of shifts, patterned or stylized shifts. Subject does not conform to the stated objectives of the task of earning the highest score, but holds to a set mode of response. Responses are not based on previous experiences of success or failure. Momentary changes in the situation are apparently overlooked.
2. **Arbitrary** - High responsiveness to the test situation, high number of shifts. Subject does not conform to the stated objectives of the task, and tends to respond by goal setting that is characterized by shifting up after failure and down after success. Responses do not conform to previous experiences of success or failure.

- 3 *Flexible* Moderate responsiveness to the test situation, average number of shifts. Subject conforms to the stated rules and framework of the task, adjusts to previous success and failure by an estimation of probability rather than by holding exactly to previous attainment
- 4 *Achievement following* High responsiveness to test situation, very high number of shifts, low conformity to rules and framework of task, sets new framework. Subject conforms exactly to previous success or failure by responding in the same direction as the immediately preceding achievement
- 5 *Conforming* High responsiveness to test situation, high number of shifts, high conformity to goal objectives by adherence to the rules and framework of the task, high adjustment to previous experience of success and failure, strongly influenced by momentary experiences

Reliability of Ratings To check the reliability of both ratings a folder was prepared including (1) a typewritten copy of the instructions and descriptions of the ratings, (2) a table of norms for the various measures previously described, and (3) a copy of the original protocols of the test for each subject. This folder was given to one of our colleagues¹ who was familiar with the general outline of level-of-aspiration testing, but who had no special knowledge of the procedure or purposes of the experiment. No coaching sessions were held. Thus the reliability ratings are based on the written instructions and the rater's interpretation of them. These ratings were compared with the ratings made by the experimenter using the same criteria.

The reliability ratings on goal-level setting gave a Pearson product moment r of .963. There were seven possible categories and complete agreement in ratings was found in 76 per cent of the cases. In the other 24 per cent of the cases the raters disagreed by no more than one category.

In getting the reliability of the method of adjusting goals to success and failure ratings, we had recourse to coefficients of contingency, since we could not presume that our categories were continuous. The coefficient of contingency was $c = .823$ and corrected for 5×5 tables (16, p. 398) $c = .920$. With five possible categories there was an 84 per cent complete agree-

¹ Grateful acknowledgment is made to Dr. Morris Roseman, clinical psychologist, VA Hospital, Lexington, Ky.

ment. The reliability of ratings was also high for this variable.

The Patterns that Emerged. In Table 5a we have the array of the experimental group as it distributed itself in light of the two ratings we used. As can be noted, achievement following had no candidates for its interval. It may also be questioned whether the use of seven step intervals is a refinement that is particularly helpful in discriminating behavior, and whether a cruder differentiation might not be more appropriate.

TABLE 5a
Distribution of Ratings: Experimental Group N = 50

Height of Goal Level Setting	Rigid	Method of Adjusting Goals to Success and Failure				Total
		Added	Fixed	Sub. Fctd	Cost	
1	0	0	2			4
2	3	2	2			7
3	2	1	1		2	6
4	2		2		2	6
5	4	4	2		1	14
6	0	6	0		2	7
7	1	2				3
Total	12	14	12	0	10	50

We therefore condensed the ratings of the height of goal-level setting by including in one interval those above 67 SD and, in another interval, those below 67 SD. The group between 1.67 and 67 SD was combined in the other interval.

By the elimination of the achievement following classification we had four methods of adjusting goals to success and failure. Thus with three levels of goal level setting and four methods of adjusting goals to success and failure we ended with twelve distinct patterns of behavior (Table 5b).

TABLE 5b
Distribution of the Experimental Group into the Table of Patterns N = 50

Height of Goal Level Setting	Rigid	Method of Adjusting Goals to Success and Failure			Total
		Added	Fixed	Cost	
67 SD and over	4	3	4	0	11
67 SD to 67 SD	8	6	7	2	23
67 SD and over	2	5	1	4	12
Total	14	14	12	10	50

The patterns may be entitled as follows:

1. High positive goal level setting and rigid adjustment

- II Medium goal level setting and rigid adjustment
- III High negative goal level setting and rigid adjustment
- IV High positive goal level setting and arbitrary adjustment
- V Medium goal level setting and arbitrary adjustment
- VI High negative goal level setting and arbitrary adjustment
- VII High positive goal level setting and flexible adjustment
- VIII Medium goal level setting and flexible adjustment
- IX High negative goal level setting and flexible adjustment
- X High positive goal level setting and conforming adjustment
- XI Medium goal level setting and conforming adjustment
- XII High negative goal level setting and conforming adjustment

Inspection of the table of patterns (Table 5b) for our experimental groups finds no subject using pattern X, one subject using pattern IX, and two using patterns III and XII. With the limited number of subjects in our experimental group, conclusions as to the acceptability of patterns cannot be made. It may, however, be noteworthy to observe that, with the exception of the arbitrary adjustment, high negative goal-level setting was not common for our group.

Positive goal level setting has indeed been reported as the more typical behavior on level of aspiration tasks in our Western culture by most experimenters (19, 14). As we have observed elsewhere (4) in a group of 25 adjusting college girls we found no negative goal level setting. But negative setting has been reported by Arluck (2) for epileptics and by Miller (15) in conversion hysteria. We have also reported negative goal setting in hypertensive patients (4). Thus, the presence of negative goal level setting in our experimental group may be attributed to the hypertensive patients within the sample.

As for the modes of adjusting goals to success and failure, we have reported elsewhere (6) that the college girl group used

the rigid, conforming, and flexible methods, and not the arbitrary. In the same study our hypertensive patients used the arbitrary and conforming methods predominantly and our asthmatic group preferred the rigid mode of adjustment.

It would thus appear that negative goal level setting and the arbitrary method of adjusting goals to success and failure, by their incidence among certain patient groups and absence from certain adjusted groups, may be suggestive of less acceptable methods of adjusting.

We may also raise a question as to the acceptability of high positive goal level setting in view of the observations of Sears, who described it as an effort to secure commendation rather than a realistic goal.

In consideration of the above, a tentative formulation of the behavior consonant with each pattern as well as some suggestion as to the motivation for such behavior might be offered in the following:

- I. Setting goals well above achievement, sticking to such goals despite failure to achieve. Subject appears to place greater value on high goal statement than on achievement, is easily threatened by potential failure; sets and sticks to high goals as a means of gaining reward for effort.
- II. Setting goals within the range of achievement, sticking to such goals despite fluctuations in achievement. Subject sets reasonable goals, appears to fear failure, to handle fear by cautious shifts in goals.
- III. Setting goals well below achievement, sticking to such goals despite easy success in attainment. Subject appears to fear failure, insures success by an extremely low bid which is persisted in in the face of frequent assurance of ability. Fear of failure seems overwhelming and is dealt with by retreat to safe but unreal position.
- IV. Setting goals well above achievement; frequent changes in goal level with retreats from success and overcompensations for failure. Subject appears to strive for the appearance of success, but each trial is a challenge that calls for emergency adjustment. Having succeeded, he fears he will not be able to do well and retreats;

having failed, he wishes to do well and overcompensates by high goal statement

- V Setting goals within the range of achievement, frequent changes in goal level with retreats from success and overcompensations for failure. Subject appears to set reasonable goals, but such goal setting seems fraught with tension for him. Having succeeded, he fears he will not be able to do as well and retreats, having failed, he wishes to do well and compensates by high goal statement
- VI Setting goals well below achievement, frequent changes in goal level with retreats from success and overcompensation for failure. Subject appears to fear failure and sets low goals to insure success, but such low goal setting and its attendant success are not satisfying and call for strong assertion of high goals. He seems so uncertain, however, that the assertion of goal is overwhelmed by the retreat from a potentially precarious position.
- VII. Setting goals well above achievement, changing goals in light of trends in achievement. Subject appears to want to do well. Apparently recognizing the unrealistic height of goal level, he uses it primarily as incentive for better performance
- VIII Setting goals within the range of achievement, changing goals in light of trends in achievement. Subject appears to keep goals within a modest range and to be able to view his own performance with some objectivity. Goal setting seems primarily a rational judgment of probability
- IX. Setting goals below achievement, changing goals in light of trends in achievement. Subject appears to be cautious and conservative in goal-level setting, yet sufficiently objective to resist momentary changes or holding rigidly to one goal.
- X. Setting goals well above achievement; changing goals in conformity with previous success and failure. Subject appears to strive for high achievement, shifting continually to take advantage of every assurance and

to retreat with every rebuff. Earnest desire to do well overcomes to some extent the fear of failure.

XI Setting goals within the range of achievement; changing goals in conformity with previous success and failure, cautiously advancing and retreating, holding as closely as possible to achievement

XII Setting goals well below achievement, changing goals in conformity with previous success and failure. Dominated by a desire to succeed, subject plays safe in goal setting. Desire to do well is overwhelmed by fear of failure

Summary

Descriptions of patterns of behavior on level-of-aspiration tasks have emphasized the height of the average goal-discrepancy score as criterion. Further observations on the manner in which goals are set have indicated the propriety of considering more than one major variable in describing behavior in level of aspiration tasks.

Using fifty subjects on the Rotter aspiration board and task and analyzing the various scores that had been previously suggested, as well as others new with this study, three major variables seemed to emerge from the table of intercorrelations. Of these, goal level setting and the method of adjusting goals to success and failure were the ones investigated. Judgment was the third variable, but its relevance was not thoroughly explored.

Rating scales were developed for the two variables in order to insure the best assessment of the test protocol. Ratings appeared to be reliable.

A condensation of the table of distribution of subjects showing the relationship of the two variables (goal-level setting and method of adjusting goals to success and failure) resulted in the denotation of twelve patterns of response. Reference to other populations to whom this scheme of patterns was applied suggested the usefulness of the method.

REFERENCES

1. Adams, D. K. "Age, Race, and Responsiveness of Levels of Aspiration to Success and Failure." *An Abstract Psychological Bulletin*, XXXVI (1939), 573

2. Arluck, E. W. "A Study of Some Personality Characteristics of Epileptics." *Archives of Psychology*, CCLXIII (1941), 77
3. Binger, C. A., Ackerman, N. W., Cohn, A. E., Schroeder, H. A. and Steele, J. M. "Personalities in Arterial Hypertension." *Psychosomatic Medicine Monograph No. 8*, 1945, 228
4. Cohen, L. D. "Level of Aspiration Behavior in Certain Psychosomatic Disorders." Ph.D. dissertation, Duke University, 1949
5. Cohen, L. D. "Methods of Adjusting to Success and Failure in Certain Chronic Medical Disorders." An Abstract *J. Elisha Mitchell Scientific Society*, LXV (1949), 198
6. Duffy, Elizabeth. "A Systematic Framework for the Description of Personality." *Journal of Abnormal (Social) Psychology*, XLIV (1949), 175-190
7. Frank, J. D. "Individual Differences in Certain Aspects of the Level of Aspiration." *American Journal of Psychology*, XLVII (1935), 119-128.
8. Gardner, J. W. "The Use of the Term Level of Aspiration." *Psychological Review*, XLVII (1940), 59-68
9. Garrett, H. E. *Statistics in Psychology and Education* New York Longmans, Green and Company, 1947
10. Hausmann, M. F. "A Test to Evaluate Some Personality Traits." *Journal of General Psychology*, IX (1933), 179-189.
11. Hoppe, Ferdinand. "Erfolg und Misserfolg." *Psychologische Forschung* XIV (1930), 1-62
12. Klugman, Samuel F. "Relation Between Performance on the Rotter Aspiration Board and Various Types of Tests." *Journal of Psychology*, XXIII (1947), 51-54
13. Klugman, Samuel F. "Emotional Stability and Level of Aspiration." XXXVIII (1948), 101-119
14. Lewin, K., Dembo, T., Festinger, L. and Sears, P. S. "Level of Aspiration." *Personality and the Behavior Disorders* (J. McV Hunt, Ed.). New York. Ronald Press, 1944 Pp 333-378
15. Miller, D. R. "Levels of Aspiration of Hysterics and Neurasthenics." An Abstract *American Psychologist*, II (1947), 406
16. Peters, C. C. and Van Voorhis, W. R. *Statistical Procedures and Their Mathematical Bases* New York. McGraw-Hill Book Co., 1940
17. Rotter, Julian B. "Level of Aspiration as a Method in the Study of Personality: I. A Critical View of Methodology." *Psychological Review*, XLIX (1942), 463-474
18. Rotter, Julian B. "Level of Aspiration as a Method of Studying Personality: II. Development and Evaluation of a Controlled Method." *Journal of Experimental Psychology*, XXXI (1942), 410-422
19. Rotter, Julian B. "Level of Aspiration as a Method of Studying Personality: III. Group Validity Studies." *Character and Personality*, XI (1943), 255-274.
20. Rotter, Julian B. "Level of Aspiration as a Method of Studying Personality: IV The Analysis of Pattern Response." *Journal of Social Psychology*, XXI (1945), 159-177

21. Sears, P. S. "Level of Aspiration in Academically Successful and Unsuccessful Children." *Journal of Abnormal (Social) Psychology*, XXXV (1940), 498-536.
22. Sears, R. R. "Level of Aspiration in Relation to Some Variables of Personality: Clinical Studies." *Journal of Social Psychology*, XLV (1941), 311-336.
23. Sears, P. S. "Success and Failure: A Study of Motility." *Studies in Personality* (Q. McNemar and M. A. Merrill, Eds.), New York: McGraw-Hill Book Co., 1942. Pp. 235-258.
24. Shakow, D., Realnick, P. and Lebeaux, T. "A Psychological Study of a Schizophrenic: Exemplification of a Method." *Journal of Abnormal (Social) Psychology*, XL (1945), 154-174.
25. Weiss, Edward and English, O. S. *Psychosomatic Medicine*. Philadelphia: Saunders, 1943.

EVALUATION OF AN OPTOMETRIC TEST

A. R. LAUER

Iowa State College

and

WILLIAM B. MICHAEL

San Jose State College

Background

VARIOUS studies have been made of professional aptitudes since Moss and others developed a test of medical aptitude at George Washington University. Although each of the several professional schools of optometry operating in the United States and Canada has used some form of entrance examinations or selective techniques to admit applicants, only one previous study concerning optometric aptitude per se has been published (11) to the knowledge of the writers.

Although during the past four years the optometric course in some institutions has been raised to five years, the schools are besieged by hundreds of applicants who cannot be admitted. Both the schools and the profession have found it advisable to select for admission only those best qualified to handle the courses offered—courses which are heavily loaded in mathematics, physics, and the basic biological sciences.

The Los Angeles College of Optometry operates on a five-year program with two preparatory years required for entrance. To fulfill these requirements, however, most students have to present nearly three years of standard collegiate credit in stipulated areas. In fact, many applicants have completed the bachelor's or master's degree, while a few have already received the Ph.D. degree. In addition, there is a sprinkling of applicants who have been graduated in law, dentistry, medicine, and osteopathy.

The Los Angeles College of Optometry normally admits approximately one hundred applicants once a year. Since the

curriculum had been completely revised into a five-year program, it was decided to admit about 150 in the First Semester of 1945. The increment in enrollment was partly justified on the basis that the College would not graduate a class in 1950 because of the introduction of an additional required year in the new course of study.

Problem

Consequently, it was deemed advisable to develop a satisfactory test for use in evaluating applicants, since pre optometry grades were only moderately prognostic in that they yielded a correlation of only $r = .312$ with grades in professional optometry. The lack of a higher degree of correlation was probably due in large measure to the presence of systematic differences in marking standards at various colleges and to differences in types of pre optometry curricula pursued by the applicants.

It was decided by the Administration to suit several hundred applications and to allow about twice as many as would be accepted to take the entrance examination. A total of 229 applicants completed all tests. A group of this size was deemed to be sufficient for undertaking an initial exploratory investigation of the reliability of the test units and for ascertaining the extent to which the measuring instrument would reflect individual differences in various abilities and traits hypothesized to be of importance in the course of study.

The general hypothesis to be investigated may be stated as follows: Success in professional optometric courses is a function of, (1) general intelligence or mental alertness, (2) general cultural background, i.e., familiarity with well-known works in the fine arts and with important individuals associated with the fine arts, (3) attitudes toward scholastic activities, and (4) background (achievement) in the basic sciences. Four test units were employed to measure these four hypothesized functions—one test for each function.

Nature of the Test Units

The tests were grouped into two sections with a general sheet of directions for the examiners. The first section was a battery of four basic-alertness or general-classification subtests, each of which was administered on a strict time-limit

basis. The second section consisted of three separate tests administered as amount-limit instruments. Descriptions in slightly more detail may be given as follows.

Section One (Form A) Test I consists of four sub-tests. Sub-Test 1 is a revised form of an arithmetic test of 10 scaled items originally used in the Army Alpha. Sub-Test 2 is a revised form of an opposites test consisting of 20 scaled items originally used in one form of the Army Alpha. Sub-Test 3 is a revised form of a cube-counting test consisting of 8 scaled items used in the Army Beta. Sub-Test 4 consists of 12 items selected from Canfield's (11) test, which has been used as an optometric entrance examination. The working time of this section was twelve and one-half minutes. The first three of these sub-tests have been described as a battery by Rostion and Lauer (7), and have been used in printed form for industrial applications.

Section Two (Form A) Test II is a form of cultural inventory standardized by Lauer and described in a previous publication (3). Originally administered as a seven-answer multiple-choice type of test having an estimated reliability of +.94, it was revised to be used as a five answer multiple-choice test in conjunction with standard IBM scoring forms.

Test III is a test of scholastic attitudes covering the reaction to such items as library study, class attendance, and laboratory work, assumed to contribute toward success in college. The general form of test has been described by Lauer (4).

Test IV is similar to Test II but covers biological material. The original form was used by Schweet and Lauer (8). It was also revised to be used as a five answer test with standard IBM scoring sheets.

The approximate working time required for Section Two is thirty minutes, although all examinees were allowed to finish the test. In order that the testing period might not be unduly prolonged, some verbal incentive such as "get through as quickly as possible," or a similar suggestion, was made.

Administration and Scoring of Tests

Most of the tests were administered by the senior author during the Spring Semester of the 1947-1948 academic year.

About 20 per cent were given on the same day in two successive periods. The other 20 per cent were given under authorized and capable supervision either at the College or by qualified personnel throughout the country.

The battery of tests was mimeographed to be used with standard IBM electric scoring sheets. As mentioned previously, the subtests of Section I were closely timed. A priori scoring formulas were thus employed to correct for chance successes on multiple choice items. For the last three units the amount-limit method of administration was used. No negative weights were assigned to any responses in the attitude test but a system of reverse positive weights was used for the negative items, in which high values were added as low values, and vice versa.

TABLE 1
Reliability Coefficients of Test Units

Test Unit	Nature of Test	N	Reliability Coefficients Estimated from Correlations
1	Attitude as Measured (Intelligence)	124	75
2	Cultural Interest	127	84
3	Technical Aptitude	120	81
4	Background Background	121	82

Statistical Procedures and Results

Reliability coefficients of the various units of the test battery were estimated from correlations of the scores of all applicants taking the tests. The split half method was used in which the odd numbered and even numbered items attempted were correlated, followed by the application of the Brown-Spearman formula. The estimated reliabilities of the test units are presented in Table 1.

A reliability coefficient was similarly estimated from correlations of measures within the criterion (first semester grades in the professional optometry course). Initially, two sets of courses were picked at random such that half the courses were in one set and half in the second set. No systematic combinations of highly related content courses resulted. The product-moment correlation obtained between grade averages in the two sets of courses after correction for length was .88 ($N = 133$),

which is probably to some extent spurious. The magnitude of the coefficient may have been due, in part, to the rôle of a possible "halo effect" in the assigning of marks in view of the small number of instructors and the degree of unintentional cooperation among them.

There were 133 subjects for whom data were available for each test variable. The product-moment intercorrelations of the tests and criterion are presented in Table 2. The multiple-correlation coefficient was calculated by Wallace and Snedecor's version (10) of the Doolittle method. In addition to this coefficient, the beta weights and the contributions of each test to the total predicted variance are presented in Table 3.

TABLE 2
Intercorrelations of Tests (1, 2, 3, 4) and Criterion (5) $N = 133$

	1	2	3	4	5
1		.441	.083	.318	.481
2			-.094	.394	.328
3				.168	.271
4					.278
5					
Means	27.57	51.60	133.49	54.53	45.04
Standard deviations	5.73	10.40	18.72	9.74	11.58

Legend: See Table 1

When the relative degree of homogeneity of the restricted group was considered, an obtained coefficient of multiple correlation of .56 appeared to be substantial. The standard deviations of scores on the various test units for the selected group of 133 students were from three-fourths to four-fifths the size of those obtained for the total (unrestricted) group of 229 applicants. On the assumption that the standard deviation of the composite score of the total number of applicants taking the tests ($N = 229$) would have been about 25 per cent larger than that of the restricted group ($N = 133$), the magnitude of the probable correlation between the composite score and the criterion grade average of the total group (had all applicants been admitted to training) might be estimated to be in the vicinity of .465.¹

¹ An improved estimate of what the obtained coefficient of multiple correlation would have been for the total sample may be found as follows: first, the correlation in the unrestricted group between each of the four test variables, upon which restriction

Examination of Table 1 reveals that the test in *Alertness* (General Intelligence) stands far above the others in predictive value. Next in importance is the test of *Scholastic Attitudes*. The advantages gained by this test are its low correlation with both the intelligence test and the other tests and its significant correlation with the criterion.

Such a finding suggests the possible value of tests of this type in selection procedures. It should, however, be tried with new samples to determine whether it will consistently contribute to the prediction of academic success. Further refinements of the test, including new item analyses and the possible addition of other items, are being undertaken with the view of increasing its validity.

The contribution of the *Cultural Inventory* to the total pre-

TABLE 1
Multiple Regression Data

Test variable	1	2	3	4
Ratio weight	72.0	16.77	24.39	.0536
Contribution to predicted variance	17.61	25.51	56.58	.0146
$R^2_{\text{total}} = .160$	$R^2_{\text{total}} = .41$ (corrected for shrinkage)			

Legend for Table 1

dicted variance was substantial. In view of its high correlation with the test of *Alertness*, it would not be expected, however, to make a large unique contribution to the multiple correlation.

The test in *Biological Background* failed to add much to the total predicted variance. Since most of the work in the First Semester tends to be more closely allied to background in mathematics and the physical sciences, this test may show its value in subsequent semesters.

General Summary and Conclusions

A battery of four test units was assembled for the evaluation of optometric aptitude. Of the 229 applicants to whom the battery was administered, 131 entered and completed the first semester's work at the Los Angeles College of Optometry.

First, the criterion may be estimated through use of a formula given by Thorndike (9, 171), since the standard deviations of test scores are available for the total group and restricted group; second, the test intercorrelations for the total group may be calculated; third, with these data both new regression weights and the multiple-correlation coefficient may be computed through use of the Bockstette method.

In terms of the results obtained from a correlational analysis of the data for the sample studied, the following conclusions may be drawn

1. The best single test used for prognosticating aptitude in professional optometric training was one of general intelligence. Although the machine-scored technique appeared to lower the reliability of the test somewhat and although the group selected tended to be relatively homogeneous in ability, a correlation of .48 of the test with the criterion was obtained.

2. The second-best test used in the study was one in scholastic attitudes. The inclination to study, as measured by the test, did not seem to be significantly correlated with intelligence. Such a finding should be further explored.

3. Cultural knowledge was quite strongly correlated with grades, but partly because of its being closely associated with intelligence.

4. Biological background itself did not seem appreciably to be associated with success in optometric training during the First Semester. This may have been due to the fact that this semester is weighted heavily with materials of a mathematical and physical nature.

5. The results obtained would seem to warrant the advisability of submitting every preoptometry student to a similar battery for at least two purposes: (a) To aid in developing the best procurement technique possible. (b) To assist in assigning responsibility for good work to individual students at the beginning of the professional course in optometry.

Suggestions for Future Research

As to the improvement of the over-all validity of a test battery for selection of students for optometric training, additional tests might be included advantageously. Tests which measure the psychological traits of spatial relations, visualization, and perceptual speed might possess differential validity (5,6). A preliminary job analysis of activities involved in various laboratory subjects of the curriculum has indicated such a possibility. A biographical data blank might also be of considerable value to over-all prediction, if a valid scoring key could be achieved.

There is a hope that the present view of comprehensive factor-analysis as a means of selecting aptitude tests and criterion measures for personnel selection would probably aid in the selection of a comprehensive number of relatively pure tests that would measure diverse factors and greater importance uncovered in the criterion. In this regard the realization of an economy in the number of tests and the simplified, relatively pure factor scores (1, 2) could be attempted for purposes of prediction of success in different subject matter areas of professional and theory courses. Such a procedure would permit a further degree of insight into a student's potentially strong, positive and weak points than would a more complete score derived from the multiple regression equation.

REFERENCES

1. Guilford, J. P. "The Factor Analysis of Test Development Procedures," *Psychological Review*, LV (1948), 9-24.
2. Guilford, J. P. and Michael, W. B. "Approaches to Univocal Factor Scores," *Psychometrika*, XVII (1948), 1-22.
3. Loeber, A. R. "Measurement of Cultural Knowledge," *Journal of Educational Psychology*, XXIX (1936), 287-91.
4. Loeber, A. R. "The Reliability of Background Interest, and Study Habits Scales in College," *Abstracts Proceedings of the Iowa Academy of Science*, XLVII (1944), 314.
5. Michael, W. B. "The Nature of Space and Visualization Abilities," *Transactions of the New York Academy of Sciences*, Series II, II (1949), 26-31.
6. Michael, W. B., Zimmerman, W. S., and Guilford, J. P. "An Investigation of Two Hypotheses Regarding the Nature of the Spatial Relations and Visualization Factors," *Educational and Psychological Measurement*, X (1950), 19-212.
7. Reardon, Charles and Loeber, A. R. "The Reliability of an Abreviated Test of Alertness," *Proceedings of the Iowa Academy of Science*, XLI (1939), 319-20.
8. Schaefer, Richard and Loeber, A. R. "Preliminary Evaluation of a Test for Biological Background" (abstract), *Proceedings of the Iowa Academy of Science*, XLVII (1949), 391.
9. Thorndike, Robert L. *Personal Selections, Test and Measurement Techniques*. New York: John Wiley, 1949.
10. Wallace, Henry A. and Snedecor, George. *Correlation and Machine Calculation* (revised edition). Iowa State College Official Publication, Vol. XXX, No. 4.
11. Warren, Neil D. and Guilford, J. P. "An Optometric Aptitude Test," *Educational and Psychological Measurement*, VIII (1948), 183-191.

THE EFFECT OF CLIENT PARTICIPATION IN 'TEST' INTERPRETATION

PAUL L. DRISSLE AND ROSS W. MARLSON

Michigan State College

The Problem

The role of tests in the counseling process is as varied as the viewpoints about counseling. At the present time, test administration and interpretation practices generally suggest a rather directive, authoritarian type of counseling. The counselor, in effect, says, "If you will unquestioningly fill out the forms and take all the tests I prescribe, I will be able to tell you what to do." The inability of counselors to break away from this concept of tests is probably one basic reason for the disinterest in tests commonly exhibited by the Rogerian adherents. At the verbal level, at least, few will quarrel with the idea that counseling should seek for the *development of the client's self understanding*, self-acceptance, and self-sufficiency, always with due regard for his social responsibility. Acceptance of this viewpoint involves the obligation of determining whether, and how, tests contribute to this development. Our attention then focuses more on the client's impressions and reactions than on the test data.

Test Interpretation Practices

A review of counselor procedure in test interpretation at the Michigan State College Counseling Center indicated wide variation in practice. *Individual counselors claimed to vary their test interpretation procedure greatly in terms of the needs and personality of the individual client.* Most counselors felt that the entire staff functioned in about the same way in dealing with this aspect of counseling. Actual practices may be listed as follows.

1. *Counselor gives an interpretation directly from test data.* The client may see the data, but it has little or no meaning to him

2. *Counselor uses a test profile, but uses it only as a visual aid in his discussion.*

3. *Counselor gives a very general summary of the test results and emphasizes the implications rather than the test data.*

4. *Counselor questions client and urges client to comment and question at all stages of interpretation.*

5. *After a brief introduction to the profile, the client is allowed to question, comment, and discuss at will.* The counselor concentrates on maintaining the client's flow of thought.

These practices differ largely in the amount of client participation involved - varying from almost a *counselor lecture* to a *client led discussion*. Several articles in the past few years have dealt with matters closely related to this.

Rogers (4) has stated that the client centered counselor makes less use of tests and uses them in a different fashion. "They (psychometric tests) do not stand up well as a technique for client centered counseling." He indicates, however, that tests may be introduced at the client's request, but that even then the focus of counseling remains on the emotional attitudes expressed. Admitting the value of tests in selection and in research, Rogers holds that, eventually, tests initiated by the counselor hinder the counseling process whose purpose is to release growth forces.

Bordin and Bixler (2) describe an interview procedure wherein the process of test selection serves to enable the client better to understand the significance of materials related to his own feelings and to facilitate the development of a deeper understanding of the problem. They hypothesize, further, that this test selection procedure serves as motivation for the testing program itself and that it fosters the client's recognition of his own responsibilities in the counseling process. These writers suggest needed research studies involving the effect of clients' participation on their acceptance of the test results, their assumption of responsibility for solving their problems, and their attitudes toward taking tests and toward the counseling process afterward. In a study of the kinds of test choices clients make and their behavior during the test selection process, Seeman (6) had fifty electrically recorded interviews classified by qualified judges as to the division of counselor-client responsibility.

The experimental counselors were found to be consistent in their use of the client self-selection method of choosing tests. It was found that clients selected tests available for prediction in 93.2 per cent of the possible cases, and that their reactions to and during the process varied.

Bixler and Bixler (1) discuss categories of test-interpretation technique involving varying degrees of counselor opinion. They conclude that counselors should:

1. Give the client simple statistical predictions based upon the test data.
2. Allow the client to evaluate the prediction as it applies to himself.
3. Remain neutral toward test data and the client's reaction.
4. Facilitate the client's self-evaluation and subsequent decisions by the use of therapeutic procedures.
5. Avoid persuasive methods. Test data should provide motivation, not the counselor.

A study of the relationship of the amount of counselee talk to the effectiveness of counseling is reported by Carnes and Robinson (3). Analyzing typewritten transcripts of 78 interviews, they found a wide range in the amount of client talk and growth in client insight, with the topic of the unit influencing the relationship. The writers conclude that, since causal relationships are not clear, "it is not possible to use the amount of client talk as a criterion of counseling effectiveness."

It will be seen that none of the above references seems to have been concerned with exactly the problem discussed here.

Specific Hypotheses for Investigation

The issues raised in the preceding discussion were formulated into three major hypotheses for investigation.

Hypothesis I Clients who participate more actively in the test-interpretation process gain more in self-understanding than do those who participate less.

Hypothesis II Clients who participate more actively in the test-interpretation process are more certain of their final vocational choice than are those who participate less.

Hypothesis III Clients who participate more actively in the test interpretation process are more satisfied with the experience than are those who participate less.

It is apparent that these hypotheses do not cover all the

described and involved the use of tests in the counseling process. The quality of test acceptance and self-efficacy are not mentioned except as they may be elements of the client's satisfaction. However, the research problem growing out of the stated hypotheses seemed sufficiently complicated for an initial investigation.

Development of Criteria and of Evaluative Instruments

Any investigation of "client participation" must be preceded by a clarification of the phrase. The following list of principles was developed and these, for the purposes of this study, define client participation:

1. The test should not be given until it is clear that the client (a) has nothing more important to him for discussion and (b) is emotionally ready to deal with them.
2. A test profile should usually be used as the most meaningful and simplest form for giving test results to a client. Ability, achievement, interest, and aptitude scores may readily be placed on one profile. Personality ratings must frequently be handled separately, partly because of their nature and partly because the interpretability of scores is differently interpreted.
3. Explain the general basis of the profile emphasizing the different qualities measured and the comparison of the individual with other individuals.
4. Avoid raising or discussing technical and statistical details as much as possible.
5. Encourage the client to voice his own explanations, hunches, and feelings.
6. Answer the client's questions but do not permit information giving to result in ignoring the emotional content of the questions.
7. Recognize the client's feelings and concerns but do not offer consolation except as it is incidental to supplying additional facts for interpretation. Above all, do not get involved in defending the tests or in impressing on the client results not yet acceptable to him.
8. Allow the client adequate time to react to the individual elements of the profile. Do not rush him into formulation of conclusions.

9. Encourage the client to relate his own experiences and other known facts to the test results

10. Do not, by silence or otherwise, seem to give tacit approval to statements which are erroneous, but do not be too hasty in correcting them

11. If the client sees only one course of action implied by the test results, do not at once suggest others, but do not permit him to judge by default that you concur.

12. The counselor should suggest the advisability of additional information, if the client does not himself sense the need. In general, it is better that occupational information be introduced on request

13. Test scores should not be too heavily emphasized since they are probably less important than a multiplicity of other factors in determining the client's final course of action

The general import of these principles is that the client is to be given the opportunity to ask questions, venture his own hunches and, in short, to develop the counseling session in the direction of his own interests and concerns. If he leads it away from test results it is assumed that there is greater value in following his stream of thought than in leading him back to the test results. Although these principles have been stated as imperatives, it is to be kept in mind that their appropriateness in test interpretation is really the hypothesis to be investigated. For the purposes of the study the counseling sessions were now to be rated in terms of the extent to which these principles were followed, so that it became necessary to develop from them a rating scale for use by judges. After several trials the scale shown in Figure I was agreed upon as the basic instrument for this purpose. While not entirely satisfactory to the judges, it was sufficiently clear that on several trials the independent judgments of interviews gave very close agreement on the separate items as well as on the total scores. The rating scale was not shown to the counselors involved in the study.

It was also necessary to develop a test of self-understanding in order to determine the actual increase in the client's self-knowledge of qualities tested. This test included questions testing the client's understanding of the level of his various abilities in relationship to norms established for various groups.

Figure 1: A Psychological Measurement

2010-2011
401-2010-2011

2010-2011
2010-2011

2010-2011
2010-2011

The following descriptions

The following descriptions

- (1) The following descriptions
- (2) The following descriptions
- (3) The following descriptions
- (4) The following descriptions
- (5) The following descriptions

The following descriptions

- (1) The following descriptions
- (2) The following descriptions
- (3) The following descriptions
- (4) The following descriptions
- (5) The following descriptions

The following descriptions

- (1) The following descriptions
- (2) The following descriptions
- (3) The following descriptions
- (4) The following descriptions
- (5) The following descriptions

The following descriptions

- (1) The following descriptions
- (2) The following descriptions
- (3) The following descriptions
- (4) The following descriptions
- (5) The following descriptions

- (1) The following descriptions
- (2) The following descriptions
- (3) The following descriptions
- (4) The following descriptions
- (5) The following descriptions

The following descriptions

- (1) The following descriptions
- (2) The following descriptions
- (3) The following descriptions
- (4) The following descriptions
- (5) The following descriptions

The following descriptions

- (1) The following descriptions
- (2) The following descriptions
- (3) The following descriptions
- (4) The following descriptions
- (5) The following descriptions

The following descriptions

- (1) The following descriptions
- (2) The following descriptions
- (3) The following descriptions
- (4) The following descriptions
- (5) The following descriptions

The following descriptions

- (1) The following descriptions
- (2) The following descriptions
- (3) The following descriptions
- (4) The following descriptions
- (5) The following descriptions

Figure 1

Appended to this for convenience in administration were questions designed to ascertain his feelings of vocational security as well as the extent of his satisfaction with the counseling

received. These two parts of the test are handled separately in the analysis. The test itself is not reproduced here because of the space it would require.

The necessity for using a common set of questions imposed the necessity of using a group of students exposed to the same testing experiences. This, in turn, imposed the necessity of some similarity in the nature of the clients involved in the investigation. It was agreed that no preference freshmen (those entering college without a declared major) would be used as subjects. The use of a common test battery with this group was appropriate for initial exploration of vocational choice. After the experimental interviews counselors were free to use additional tests if desired.

For the purpose of this study the uniform test battery included

- 1 *A.C.P. Psychological Examination*
- 2 *Cooperative Reading Test*, (Higher level).
- 3 *Minnesota Paper Form Board*
- 4 *Minnesota Clerical Test*
- 5 *Kuder Preference Record*
- 6 *California Test of Personality*.

A special test profile chart was prepared for use of counselors, but it was in no sense unique and hence is not included here.

The Experimental Design

The experimental procedure involved the following steps:

1 Development of the materials presented or described above.

2 Selection of a number of counselors interested and willing to cooperate in the study. (Originally it was intended that seven counselors would handle seven clients each, but illness, leaves of absence, and other unexpected factors resulted in unequal numbers of clients for each counselor.)

3 Selection by the cooperating counselors of the subjects for the study. The client's willingness to cooperate and to have the interview recorded was determined before including him in the study.

4 Administration of the *Test of Self-Understanding* to each client before taking the tests of the counseling battery.

- c Administration of the uniform test battery
- d Interpretation of the test results under the same physical conditions with a wise weighing made of the entire interview
- e Rating independently by four judges of the recorded interview

* Re-administration of the *Test of Self Understanding* and completion of the appended questions dealing with satisfaction and vocational security

g Re-test in two months later of the test and questionnaire to determine retention and continued satisfaction and security

i Statistical analysis utilizing analysis of variance and covariance to determine the extent and nature of the variations.

There exist a few other points of procedure not adequately covered above but a knowledge of which is necessary for following the rest of this report. The forty clients involved were all freshmen with the same stated problems, they took the same tests and had an interpretation interview of approximately forty to sixty minutes in length. While age, sex, scholastic status and other factors were uncontrolled there was no evidence of any other uncontrolled factors. The *Test of Self Understanding* was scored by comparing the client's responses with his test results. The gains difference between pre- and post test were used to measure the increase in self understanding.

The rating of a particular interview by a judge was obtained by totalling the numbers in front of the descriptions checked by the judge. This permitted a maximum of forty points (up to five points for each of eight items). Each judge was allowed up to five 'bonus' points to be used if an interview seemed better in respect to student participation than the standard scale indicated. These extra points were little used and the scale remained essentially a forty point range.

Statistical Treatment

As a preliminary step the rating scale itself was analyzed by tabulating the individual ratings of each of the judges for each interview on each of the items covered. This analysis, when compared with corresponding gains evidenced in clients' self-understanding, showed that the order of the numbered

responses corresponded to larger gains in self-understanding. This was not regarded as a validation but as a check on the consistency of the weighting of the item alternatives and also as an indication that some relationship existed between participation and gain.

The individual ratings of the four judges varied for the forty interviews from a low of nineteen points to a maximum of forty. While differences were found among the judges, the agreement was very close. When the individual judge's rating of a particular interview was compared with the mean for the four judges, one half of the ratings were found to be within one point of the mean for that particular client. Application of analysis of variance to the ratings gave the results shown in Table 1. While the variation among raters for the same coun-

TABLE 1
Analysis of Variance Applied to Ratings

Source of Variation	Degree of Freedom	Sum of Squares	Mean Square	F
Total	159	3495.84		
Counselors	6	1881.00	313.67	70.33*
Students for same counselor	33	883.59	26.78	6.00*
Raters for same counselor	21	288.55	13.74	3.08*
Remainder	99	441.70	4.46	

* Significant at the 1% level.

selor was highly significant, its magnitude was less than that among students for the same counselor and much less than that among counselors. Differences of opinion concerning one of the interviews of one counselor accounted for the significance of the difference among raters. It was considered that the uniformity for rating was such that the sum of the four ratings would adequately characterize the participation factor of an interview.

Analysis of variance applied to raw gains in self-understanding (difference between pre- and post-test scores) showed highly significant differences in the gains made by the clients handled by different counselors. Having, also, found highly significant differences among counselors in the extent of client participation, an analysis of covariance was made to ascertain to what extent and in what way the two factors might be related.

The initial step in the analysis of covariance was computation of Table 2 (a regression client participation rating, y regression gain in self-understanding).

Since the *t*-ratio shown in Table 2 is non-significant, we are led to the conclusion that client participation does account for a portion of the differences in gains observed. This conclusion results from the fact that although the gains differed significantly from counselor to counselor, the adjustment of these by use of the participation index results in non-significance.

Further consideration of the covariance analysis involving tests of the significance of the "within" regression, of the "among" regression, of differences between the regressions,

TABLE 2
Analysis of Covariance

Source of Variations	Within-Counselor and Regression				Among Counselors		
	SS	df	MS	SS	df	MS	Mean Sq
Total	22	10	2.2	22	10	2.2	
Counselors	6	4	1.5	16	6	2.7	
Error	16	6	2.7	6	4	1.5	3.05
					10	2.2	2.2

$$F = \frac{2.7}{1.5} = 1.8 \text{ (not significant)}$$

and of the corresponding correlations suggested the following additional interpretations:

1. The within-regression ($\bar{r} = .12$) and accompanying correlation ($r = .4$) were not significant.

2. While the relationship of participation to gain for the various individual counselors varied greatly, the results were not statistically significant because of the small number of cases involved for each counselor. For example, for one counselor the correlation between gain and participation was +.66 while for another it was -.51.

3. The "among" means regression (.9877) and correlation (R_4) were found to be significant.

4. The gain of individuals assigned to a given counselor follows a different trend than the mean gains for all clients of

¹ The analysis used here followed very closely the procedures outlined by Snedecor in his *Statistical Methods*, pp. 318-324.

a counselor. In fact, the counselor means have a definite trend, the individual gains do not.

The findings of this portion of the statistical analysis indicate that the differences in the gains in self-understanding made by clients assigned to different counselors are correlated with the differences in the amount of client participation attained by the counselor. While each counselor did vary from client to client in the amount of participation, this variation was small compared to the differences among counselors and, furthermore, it did not show any consistent relationship with the client's gain in self-understanding.

A similar statistical analysis was made, using gains from pre-tests to tests given two months after counseling. These gains were considered as of more permanent nature. All significance ratios were found to be in the same relationship as those computed in the first analysis, using the gains immediately after counseling, but they all missed significance by varying amounts. The correlation between counselor means on gain and participation dropped to .63 and was below the 5 per cent level of significance.

One question in the *Inventory of Self-Understanding* asked the student to indicate his degree of vocational certainty as: confused, a bit uncertain, fairly certain, certain and secure. These responses were scored as 0, 1, 2, 3, and gains in certainty computed between pre-tests and each of the post-tests. The participation ratings and these gains were then subjected to covariance analysis in a manner similar to that mentioned previously. Non significant ratios, regressions, and correlations were found for gains from pre-test to testing immediately after counseling. The correlation between the mean participation ratings of the counselors and the gain in security was .65, which is not significant for five degrees of freedom. However, when gains from pre-test to final testing two months after counseling were used, the corresponding correlation was found to be .84, which is significant at the 5 per cent level.

The amount of satisfaction of the client with counseling was sampled by several questions which were converted into numerical scores. This reaction obviously could not be put in terms of gain from pre-test, but satisfaction scores were ob-

repeated measures study after a single client and two months later. Neither of these studies nor our studies showed any relationship between the amount of client participation and the amount of self-understanding. This is in contrast with the parallel findings of other studies.

The repeated measures and average correlations discussed are collected in Table 2. These are significant at the 5 per cent level are reported. None was significant at the one per cent level.

TABLE 2
Summary of correlations of self-understanding with client participation

	Correlation of self-understanding with client participation		Correlation of self-understanding with counselor participation	
	Pre-test to post-test	Pre-test to follow-up	Pre-test to post-test	Pre-test to follow-up
Self-understanding	0.48*	0.41*	0.38*	0.36*
Self-understanding	0.31	0.44*	0.31	0.36*
Self-understanding	0.0	0	0.26	0.36

Conclusions

Any conclusions that can be drawn from these results must be considered as tentative because of the small number of clients and counselors involved, and because of the acknowledged validity of some of the instruments used. It is also to be kept in mind that the degree of testing was restricted and that the counseling was almost entirely of the non-directive type. In full awareness of these and other weaknesses we venture several conclusions which we hope will be critically examined by others and tested in a variety of situations.

Our hypotheses that students who participate most gain most in self-understanding seems to be confirmed, but in a way somewhat different than expected. An individual counselor varies from almost no change in the amount of participation elicited but their variations seem to have little or no relationship with the client changes in self-understanding. Counselors vary greatly among themselves in the amount of client participation elicited and the mean gains in self-understanding made by their clients appear to be rather closely related to the mean client participation index. Rogers (5), in speaking

of client centered counseling, points out that it is more a matter of "attitude" than of "method." He contends that the client is apt to discern that the counselor is using a method or intellectually chosen tool and may, accordingly, not react well to it. Perhaps the individual counselors' variations in client participation are in the same class. Another possibility is that variations from the counselor's general level of client participation are forced on the counselor by the idiosyncrasies of the client.

2. The hypothesis that students who participate most are more secure in their vocational choice is given some confirmation by the data. Evidence immediately after counseling is favorable to this hypothesis and both practical and statistical significance are attached to the fact that after a lapse of two months the relationship is even greater. If this should be the case generally, it would suggest that high client-participation counselors are more successful in stimulating client growth than are other counselors.

3. The hypothesis that students who participate most are more satisfied is not confirmed by the data. Neither is there any evidence of less satisfaction.

The relationship of satisfaction to gain in self-understanding and to gain in feeling of vocational security was of the same order as that with client participation. The implication is that client satisfaction is not a very direct or reliable indicator of the effectiveness of counseling.

As already indicated, these findings are presented with the feeling that the issues raised and the highly tentative conclusions reached deserve far more extensive study.

REFERENCES

1. Bixler, Ray H. and Bixler, Virginia H. "Test Interpretation in Vocational Counseling." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, VI (1946), 145-155.
2. Bixler, Edward S. and Bixler, Ray H. "Test Selection: A Process of Counseling." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, VI (1946), 161-173.
3. Carter, Earl L. and Robinson, Francis P. "The Role of Client Talk in the Counseling Interview." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, VIII (1948), 635-644.

766 EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT

- d. Report of C. L. R. Psychological Tests and Client-Centered Counseling. *Educational and Psychological Measurement*, VI (1947), 119-122.
- e. Report of C. L. R. The Attitude and Orientation of the Counselor in Client-Centered Therapy. *Journal of Consulting Psychology*, XIII (1947), 82-84.
- f. Norman, Arthur. A Study of Client Self-Selection of Tests in Vocational Counseling. *Educational and Psychological Measurement*, VIII (1949), 227-246.
- g. Spence, George W. *Statistical Methods*. Ames, Iowa: The Iowa State College Press, 1946.

AN EXPERIMENT IN THE RATING OF ESSAY-TYPE EXAMINATION QUESTIONS BY COLLEGE STUDENTS¹

ALBERT ELLIS

The Diagnostic Center, Menlo Park, N. J.

While it is a common procedure for college (and other) instructors to have their students mark their own objective-type examination papers, the marking by students of their own essay type examinations is a much more unusual procedure. The present author, having a small enrollment in his Rutgers University class in General Psychology, decided to experiment with one form of self marking essay-type examination. The results of this experiment are reported below.

The procedure of the experiment was as follows. Eleven members of a class in General Psychology were given a surprise examination of the regular essay type during one of the class sessions. There were five essay questions on the examination, four of which were assigned a maximum score of five points, and one of which (Question No. 3) was assigned a maximum score of ten points.

The papers were collected immediately after the thirty-minute examination period had ended, and each paper was anonymously assigned a code number by the instructor. Then, one by one, the examination questions were discussed by the instructor and the members of the class, and the correct or ideal answers to each question were agreed upon. After the correct answer to each question was brought out in the course of the classroom discussion, the students' individual written answers to this question were read aloud by the instructor to the class, and each student was asked to rate each answer except his own. Thus, for each of the five essay-type questions read aloud to the class, ten ratings were given by the class members—none of whom could presumably identify any of the answers except

¹ This paper was read in manuscript by Robert M. Beechley, to whom the author wishes to express his thanks for some valuable suggestions.

his own. At the same time, the instructor also rated each answer.

When the 55 individual ratings to the questions (ten students each rating five questions) had been thus obtained, the students were asked to turn in their rating sheets, with their signatures on each sheet. The mean rating given by the students to each question answered by every other member of the class was then calculated and compared to the rating given to the same question by the instructor. A comparison of the mean ratings of the eleven students and the ratings of the instructor to the same questions answered by identical students is shown in Table 1.

From an examination of Table 1 the following observations may be made:

1. In most instances, the mean rating of the students was remarkably close to the rating given by the instructor to a given question answered by a certain student.

2. The differences between the mean ratings of the students and the instructor's ratings were almost never very consistent. In the case of the examination paper of student No. 1 in Table 1, the students consistently rated the answers to this paper a little higher than did the instructor. But in all the other cases, no such bias is apparent.

3. Considering the students' and the instructor's ratings to all the pupils' answers to each of the five questions, it would appear that in Questions 1, 4, and 5 the two sets of ratings were remarkably similar, while in Question 2 there was a fairly consistent tendency of the students to give higher ratings than the instructor, and in Question 3 a fairly consistent tendency of the students to give lower ratings than the instructor.

Rank order correlation was computed between the marks (obtained from the Total column in Table 1) finally assigned to each examination paper by (a) the means of the students' ratings and (b) the instructor's ratings. Rho was found to be .96. The Pearsonian coefficient of correlation was computed between the same sets of final marks and was found to be .91. The Pearsonian correlation coefficient was also computed between the total mean ratings given to each question by the eleven students and the total mean ratings given to each

RATING OF PLY TYPE EXAMINATION

709

項目	品名	単位	数量	金額	備考
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90

$$f(x) = \frac{1}{x^2} \quad x > 0$$

question by the instructor (obtained from the last row in Table 1). This correlation was found to be .98.

Each of the rating sheets handed in by students was then examined to see whether there was any significant relationship between the total ratings given by each student to all the others' examination papers and the final rating given to his own paper by the other students and the instructor. A Pearsonian correlation coefficient of $-.15$ was found between the ratings given by the students to others and the ratings they received from the others. This means that there was apparently a slight, but highly unreliable, tendency for the students who did best on the examination to give stricter and less generous ratings to the other students.

From the foregoing data and calculations of this study, the following conclusions seem to be warranted:

1. The mean ratings of the students to their own essay-type examination papers were very similar to the ratings given to their papers by the instructor.
2. Neither the students nor the instructor utilized consistent halo or under-rating effects in their markings of the examination papers.

The main advantages of having the members of a class rate their own essay-type examination papers seem to be these: (a) The instructor has a check on the accuracy and fairness of his own marking system. (b) The marking of the papers itself becomes a stimulating learning process which may have considerable value. (c) There seems to be virtually no come-back on the part of the students or claims of unfair marking which are so frequent after an essay-type examination which is exclusively marked by the instructor.

The main disadvantages of the student rating of essay-type examinations seem to be these: (a) The actual marking, from the instructor's standpoint, takes somewhat longer than the usual kind of instructor marking, since the answers have to be read aloud to the class, and calculation of mean student ratings must be done. (b) The reading aloud to the class of essay-type questions is quite impractical if there are many examination questions or if there are more than fifteen or twenty students in the class.

In view of the data concerning the student marking of essay-type examinations presented in this paper, and in view of the general advantages of essay-type tests which have recently been pointed out by Freeman (1), Luchins (2), Vallance (3), White (4), and others, it would seem that further experimentation with group marking is well warranted where the instructor has a fairly small class

REFERENCES

1. Freeman, F. N. "The Monopoly of Objective Tests " *High Points*, XXVIII (1946), 7-15
2. Luchins, A. S., & Luchins, E. H. "Towards Intrinsic Methods in Testing " *Journal of Educational Psychology*, XXXVII (1946), 142-148
3. Vallance, Theodore R. "A Comparison of Essay and Objective Examinations as Learning Experiences " *Journal of Educational Research*, XLI (1947), 279-288.
4. White, H. "Types of Examinations: A Compromise " *Social Education*, VIII (1944), 125-126

RELATION OF CYNICISM TO CERTAIN STUDENT CHARACTERISTICS

CHARLES O. NEFIDI and MARTIN F. FRITZ

Iowa State College

EXPRESSIONS indicative of cynicism are frequently encountered, but this characteristic of personality has been subjected to relatively little investigation. During the development of a test designed to measure cynicism, 400 college students furnished personal data as to their sex, age, religious preference, political preference, marital status, educational class level, and father's occupation. The statistical technique, analysis of variance, was used to determine the relation between the cynicism scores and the personal characteristics of these students.

The testing instrument consisted of 200 items, each concerned with a situation toward which a subject could express cynicism. A detailed description will not be given here as this is available in other publications (1, 5, 6). The following statements illustrate the types of items used: "I would say that perhaps as much as half of our tax money finds its way into the hands of grafters" and "I believe that at least 90 per cent of the girls would rather marry a poor boy whom they love than a rich man whom they do not love." A total weighted score for each subject was obtained by a simple procedure which was found on the basis of statistical study to be the optimal scoring plan (6). Attention has been given to the opposite or what might be called "idealistic" responses but these investigations will not be considered in this report (2). Copies of the test were distributed at random intervals to 400 students enrolled in psychology courses at Iowa State College. Each subject was requested to complete a short questionnaire giving certain personal data. The tests were completed and returned at the convenience of the students, without signature, in the hope that this method would produce more candid or uninhibited replies.

The scores of 387 subjects were classified by age and sex as shown in Table 1. Thirteen subjects did not indicate their age. To determine the relation of cynicism to age and sex, the data were treated by analysis of variance and adjusted for disproportionality following the method suggested by Snedecor (7). Differences significant at the one per cent level of confidence were found among the age groups (Table 2). Differences significant far beyond the one per cent level of confidence were revealed between sexes (Table 2), males exhibiting more cynicism than females (Table 1). However, there seems to be no "joint effect" of age and sex since the F -value for "interaction" is not significant.

TABLE 1
Mean Score and Frequency by Age and Sex

Age	Male		Female	
	N	Mean Score	N	Mean Score
17	11	91.08	18	80.67
18	19	91.90	73	80.58
19	10	91.50	82	69.89
20	14	92.86	42	68.00
21	11	96.85	26	88.15
22 over	5	95.19	20	84.80
Total	126	96.27	261	79.85

When significant relationships between characteristics and scores are found, it is desirable in making further analyses to control these characteristics in the classification. On the other hand, even though significant relationships are found, consideration must be given the size of each class in order to avoid emphasizing relationships based on suspiciously small numbers of cases. Although cynicism was found to be significantly related to age and sex as shown in Table 2, inspection of Table 1, giving the classification of 126 males and 261 females, shows that size of the age groups dropped to as small as ten in number. It was considered desirable, therefore, to control only on sex for further statistical treatment.

The scores of the 400 subjects were classified by sex and (1) religious preference, (2) political preference, (3) marital status, (4) educational level, and (5) father's occupation. The scores of two Jewish subjects, seven subjects expressing a pref-

erence for the Socialist party, thirteen subjects designating their marital status as either "widowed" or "divorced" or designating no marital status, twelve "special" or "graduate" students, and twenty-six subjects indicating "deceased" or making no response for father's occupation were disregarded for computational purposes. The mean scores and frequencies for the various groups are shown in Table 3. These data were treated in the same manner as described in the interpretation of Tables 1 and 2.

Religious Preference.—Students expressing no religious preference revealed the largest amount of cynicism, and Catholic subjects revealed the smallest amount among the three groups for each sex. It should be noted that the non-preference groups constitute the only comparison shown in Table 3 where the

TABLE 2
Analysis of Variance of Age and Sex

Source of Variation	df	Sum of Squares		Mean Square	F
		Unadjusted	Adjusted		
Age . . .	5	23,858.81	10,930.70	2,186.14	2.30*
Sex	1	28,630.00	15,701.89	15,701.89	16.51*
Interaction of Age x Sex	5	10,652.94	2,749.11	549.82	.58
Unexplained	375	356,632.13	356,632.13	951.02	
Total	386	398,458.00			

* Significant at the 1% level.

female mean cynical score exceeds that of the male. The differences among the religious preference groups were found to be significant at the one per cent level of confidence.

Political Preference.—The three male political groups exhibited more variability of cynicism than the three female political groups. Also, in every political preference group the males were more cynical than the females, and this sex difference was found to be statistically highly significant.

Marital Status.—In every marital status group the females were less cynical than the males. An extremely high mean cynical response characterized the engaged male group. A somewhat lower mean score was found for the unmarried (and unengaged) male group. This is just the opposite of the condition found for the corresponding female groups. Although differences significant at the one per cent level of confidence

existed among the marital status groups, the interaction of sex and marital status was not significant.

Educational Level ---Some variability was found among the educational class levels, but the differences revealed were not significant.

Father's Occupation ---Although a greater difference in cynicism was found between female farmer and non-farmer groups than between the corresponding male groups, the differences between the occupational groups were not significant.

TABLE 3
Mean Score and Frequency of Student Characteristic

Characteristic	Male		Female	
	N	Mean Score	N	Mean Score
Religious Preference				
Protestant	107	94.29	238	75.68
Catholic	13	83.29	21	65.05
None	10	104.70	9	118.56
Political Preference				
Republican	52	96.10	139	76.51
Democrat	45	89.44	71	76.06
None	29	101.03	57	72.82
Marital Status				
Married	22	95.14	10	71.90
Unmarried	89	92.12	217	77.11
Engaged	13	101.00	36	72.00
Educational Level				
Freshman	98	96.51	107	78.66
Sophomore	12	80.58	102	74.61
Junior	10	98.20	42	67.98
Senior	5	100.40	12	88.00
Father's Occupation				
Farmer	46	95.39	82	72.96
Non farmer	75	94.09	171	78.69

The above findings confirm an earlier study involving the use of correlational and chi-square techniques (4).

A summary of the F-values for the student characteristics investigated is shown in Table 4. It was concluded that after controlling on sex, highly significant differences regarding cynicism existed among the age, religious preference, political preference, and marital status groups, but the F-values for the interactions of these groups with sex were not significant in any instance. There were no significant differences after controlling on sex among the educational levels or between the two occupational groups, farmer versus non-farmer. The

F-values for the interactions of the educational levels and the occupational groups with sex were not significant. After controlling on sex, the highest F value was found for religious preference followed in order by marital status, political preference, age, educational level, and father's occupation.

It is of interest to note that in all classifications significant differences were found between the sexes, and the consistency of this finding might well call for further comment. First, it is possible that the difference is a true difference, that men really are more cynical than women. If correct, this would suggest an investigation into the nature of experiences whereby men are caused to take on a more intensely cynical attitude than women.

TABLE 4
F-Values for Student Characteristics

Characteristic	Sex Constant	Sex with Characteristic Constant	Sex Interaction
Age	2.20*	16.51*	.58
Religious Pref	3.62*	22.78*	1.30
Political Pref	4.46*	11.17*	1.32
Marital Status	6.77*	22.09*	.97
Educational Level	1.84	22.15*	.96
Father's Occupation	1.11	28.91*	1.01

* Significant at the one per cent level of confidence.

A second possible explanation would involve a criticism of the test itself, that the difference is a function of the items employed. However, an inspection of the situational statements does not seem to indicate any great likelihood of "masculine loading" which might force a greater male cynical score. Moreover, in the development of a short form of the test, it was found that the 65 most differentiating items tended to be those of a somewhat philosophical nature and therefore not particularly "sex-linked" as might be expected from more highly specific situations (3). It could be held as a third explanation that women are more inhibited and not so likely to give a strong cynical response. According to this hypothesis women could equal or even exceed men in cynicism but that the test would fail to elicit an expression of its strength. In criticism, it might be said that such a sex difference in test *attitude* is not now accepted and, were it true, would invalidate much of our present-day testing. It may also be pointed out that the practical effect of suppressing cynicism might still be the

same as a true difference. In other words, suppression of cynicism may have the same effect upon behavior as lack of cynicism.

Summary

The scores of 400 college students on a test of cynicism were treated by analysis of variance to determine the significance of the relation of cynicism to certain personal characteristics. In all classifications, males were found to be significantly more cynical than females. Variations in age groups were found, older students, in general, being more cynical than younger students. Those expressing no religious preference were most cynical, Catholic subjects the least, with Protestants occupying an intermediate position. Classified politically (Republican, Democrat, or no preference) males not only showed greater variability but in all three groups were much more cynical than females. In every marital status group (married, unmarried, engaged) males were distinctly more cynical than females with the highest mean value for the engaged group. Although differences were found in mean cynical scores when classified by educational level (Freshman, Sophomore, Junior, Senior), these variations were not significant. Father's occupation classified as farmer versus non farmer failed to reveal any significant differences in terms of degree of cynicism.

REFERENCES

1. Fritz, Martin P. "A Test Study of Cynicism and Idealism" *Proceedings of the Iowa Academy of Science*, LIII (1946), 269-272.
2. Fritz, Martin P. "Covariation of Cynicism and Idealism" *Proceedings of the Iowa Academy of Science*, LIV (1947), 231-234.
3. Fritz, Martin P. "A Short-Form Test of Cynicism" *Proceedings of the Iowa Academy of Science*, LV (1948), 319-322.
4. Neidt, Charles O. "Relation of Cynicism to Certain Other Variables" *Proceedings of the Iowa Academy of Science*, LIII (1946), 277-283.
5. Neidt, Charles O. *Analysis of College Student Reaction to the Fritz Test of Cynicism*. Unpublished M. S. thesis, Iowa State College, 1947.
6. Neidt, Charles O. "Selection of the Optimal Scoring Plan for the Fritz Test of Cynicism." *Proceedings of the Iowa Academy of Science*, LIV (1947), 253-262.
7. Snedecor, G. W. *Statistical Methods*. Ames, Iowa: Iowa State College Press, 1946. P. 284.

CHANGE IN TEACHER-PUPIL ATTITUDES RELATED TO TRAINING AND EXPERIENCE¹

ROBERT CALLIS
University of Missouri

IN view of our present knowledge, it is logical to suggest that teaching ability is so complex that it cannot be investigated efficiently as a unit. However, there are many aspects of teaching ability which can be isolated and studied independently. One of these is the aspect of teacher-pupil relations. Investigations to date indicate that teacher-pupil relations may be predicted from knowledge of teacher-pupil attitudes; however, little is known about the effect of training and experience on these attitudes. Therefore, the major problem of this study was to investigate the change that occurs in teacher-pupil attitudes during teacher-training and early teaching experience.

To measure teacher-pupil attitudes the *Teacher Attitude Inventory*, a slight extension of the one constructed by Leeds (3, 9, 10), was used. Leeds found that his Inventory would predict teacher-pupil relations reasonably well ($r = 0.60$ between Inventory scores and a multiple criterion of teacher-pupil relations). Scores on Leeds' Inventory and the one used here correlated 0.95, therefore the inventories can be considered approximately equal in validity and sufficiently valid to justify further investigation (4).

Briefly, the rationale of the *Teacher Attitude Inventory* is as follows. The inter-personal relationships between teacher and pupils are an integral part of the complex of teaching which bears directly on the mental hygiene of the classroom. It is these inter-personal relationships that we are trying to predict from a knowledge of the teacher's attitudes toward the status of children and classroom situations involving discipline and other social factors. In fact, these attitudes might also be termed

¹ This paper is a summary of a Ph D thesis of the same title on file in the University of Minnesota Library (2). The writer expresses sincere appreciation to Dr Walter W. Cook who served as major adviser to the study.

a part of a teacher's philosophy of education. Leeds' approach to the prediction problem was an empirical one. The present study was designed to determine in a general way the stability of the attitudes being measured.

Many personality inventories similar to the *Teacher Attitude Inventory* are susceptible to attempts to fake good or bad (1,5, 6,7,8,11,12,13,14). If the *Teacher Attitude Inventory* were susceptible to faking to a high degree, any change in TAI-Scores² during training and experience might be "contaminated" by intentional faking. Before the major problem was attacked, the fakability of the *Teacher Attitude Inventory* was investigated so that proper interpretation of any change in TAI-scores during training and experience could be made.

The Procedure

For the investigation of the fakability of the Inventory and the change in teacher-pupil attitudes during training and experience, six testing sequences were set up. Each sequence was composed of two testings of the same group of subjects. These sequences were.

Sequence 1—the first faking sequence, consisting of a random sample of First-Quarter juniors in the College of Education³, Fall 1947, who were first tested by standard instructions and again four to six weeks later by instructions to fake good.⁴

Sequence 2—the second faking sequence, consisting of a group of First-Quarter juniors in the College of Education, Spring 1948, who were tested first by instructions to fake good, and again, ten days later by standard instructions.

Sequence 3—the control sequence, consisting of a group of First-Quarter juniors in the College of Education, Winter 1948, who were tested and retested by standard instructions at a week to ten day interval.

70-71 sequence—the first change-in-attitude sequence, consisting of a random sample of First-Quarter juniors in the College of Education, Fall 1947, who were first tested (standard instructions) at the beginning of the school year and again six months later.

80-81 sequence—the second change-in-attitude sequence, consisting of a group of First-Quarter seniors in the College of

² Scores on the *Teacher Attitude Inventory* are referred to as TAI-scores.

³ The College of Education referred to here is the College of Education, University of Minnesota.

⁴ The subjects were instructed to attempt to make as high a score as possible; that is, answer the items as they thought a good teacher would.

Education, Fall 1947, who were first tested (standard instructions) at the beginning of the school year and again six months later.

To-TI sequence—the third change-in-attitude sequence, consisting of a group of beginning teachers who graduated from the College of Education, Spring or Summer 1947, who were first tested (standard instructions) as they graduated and again after they had been teaching for six months.

The data for the experimental sequences were compared with the data for the control sequence to determine what change in TAI scores had occurred as a result of (1) attempts to fake good, (2) teacher-training, and (3) teaching experience. Also, an analysis of the responses to individual items was made to determine which attitudes (items) were affected by training and which by experience.

The chance-half and test-retest reliability of the inventory was determined. Also, the relation of TAI-scores to a measure of intelligence was determined.

Findings

Major findings—1. The susceptibility of the Inventory to attempts to fake good was investigated by comparing the data from sequence 1 and 2 with the data of the control sequence. It was found that when the subjects answered the Inventory first according to standard instructions and second according to instructions to fake good, an insignificant increase in TAI-scores occurred ($P > .05$). (See Tables 1 and 2.) However, when the subjects answered the Inventory first according to instructions to fake good and second by standard instructions, a mean decrease in TAI scores was observed that approached significance ($P = .02$). These data suggested that the Inventory may be susceptible to faking to a limited extent so an attempt was made to construct a scale which would identify faking. After several trials, a scale was developed which showed promise of being useful in detecting faking but was not considered useful in its present form. Since it was found that the attempts to fake good produced only limited changes in TAI-scores and since there would be little motivation for the subjects in the change-in-attitude sequences to fake any more than was characteristic of the individuals, no allowances for faking were made in the analysis of the change in attitude data.

2 When the data of the Jo-JI sequence (juniors) were compared with the data of the control sequence, a significant increase in TAI-scores was observed ($P < 0.1$) (See Tables 3 and 4) This increase may be interpreted as a shift in the direction of more desirable teacher-pupil attitudes Presumably the most pertinent experiences of the subjects in the Jo-JI sequence were general courses in education which were the subjects' first experience with professional course work The correlation coefficient

TABLE 1
A Statistical Comparison of Testing Sequence 1, 2 and 3

Sequence	N	t	S.D.	Mean
(1) Standard	78	61	21.26	141.53
Taking	78		1.44 = 1.15 $P > 0.5$	141.13 $P < 0.1$
(2) Taking	44	78	17.81	117.16
Standard	44		1.44 = 1.19 $P > 0.5$	115.34 $P > 0.5$
(3) Test	57	84	16.76	135.77
Retest	57		2.06	139.96 $P < 0.1$

TABLE 2
Test of Significance of Difference in Mean Gain in TAI Scores for Sequence 1 and 2 Using Sequence 3 as a Control

Sequence	N	S.D. of gain	Mean gain
1	78	2.52	9.60
2	44	1.54	4.19
3	57	1.54	4.19
1	78	2.52	9.60
2	44	1.54	4.19
3	57	1.54	4.19

* Behrens Fisher Test of Significance between means when the variances are unequal

cient between the TAI-scores of the first and second testing in this sequence was 0.71

3 The comparison of the data for the So-SI sequence (seniors) with the data for the control sequence indicated that no significant change in TAI scores had occurred ($P > 0.5$). Presumably the most pertinent experiences of the subjects in this sequence were student teaching and courses in teaching methods The correlation coefficient between TAI-scores of the first and second testing of the So-SI sequence was 0.74.

4 The comparison of the data for the To-TI sequence (teachers) with the data for the control sequence indicated that a significant decrease in TAI scores had occurred ($P < .01$), which

TABLE 3
*Tests of the Significance of the Difference Between Mean TAI Scores in the Jo-JI, So-SI, and To-TI and Control (Test-Retest) Sequences**

Group	N	t	S.D.	Mean
Jo	175	71	21.42	140.11
JI	175		18.77	151.37
So	147	74	17.49	148.00
SI	147		10.04	151.19
To	137	69	18.51	147.45
TI	137		21.31	141.51
Test	57	84	16.76	135.77
Retest	57		20.06	139.96

* The normality of the distribution of TAI scores in the Jo-JI, So-SI, and To-TI sequences was tested and it was found that distributions did not depart significantly from normality. The test-retest sample was too small to test the normality of that distribution of TAI scores.

TABLE 4
Tests of Significance of Differences in Mean Gain in TAI Scores for Jo-JI, So-SI and To-TI Sequences using Sequence 2 (Test-Retest) as a Control

Sequence	N	S.D. of Gain	Mean Gain
Jo-JI	175	15.61	11.45
Control	57	11.54	4.19
So-SI	147	13.19	1.19
Control	57	11.54	4.19
To-TI	137	11.65	-3.94
Control	57	11.54	4.19

* Behrens-Fisher Test of significance between means when the variances are unequal.

may be interpreted as a shift in the direction of less desirable teacher-pupil attitudes. Full-time teaching was the experience of the subjects in this sequence which was presumably most pertinent to TAI-scores. The TAI-scores on the two testings in the To-TI sequence correlated 0.66.

5. An analysis of the effect of training and early experience on each of the 239 teacher-pupil attitudes (items) in the Inventory revealed that a majority of the attitudes were not affected significantly by training or experience. The first six months of professional training produced significant changes in the desirable direction in 20 per cent of the attitudes (items), while the first six months of experience produced significant changes in the undesirable direction in 11 per cent of the attitudes (items). There were only four attitudes (items) which were affected significantly both by the first six months of training and the first six months of experience (see Table 5).

TABLE 5
The Number of Items Which Showed Significant Changes (at the 5 per cent level) in the Per Cent of the Various Groups Answering the Items Correctly During the Test Retest, To-Jl, So-Sl, and To-Tl Sequences†

Direction of Change	Test Retest		Sequence					
			To-Jl		So-Sl		To-Tl	
	I	%	I	%	I	%	I	%
Increase	9	4	48	20	9	4	6	3
Decrease	6	2	6	2	1	1	27	11

* Per cent of total number of items in the Inventory.
† About half of the items which showed significant response change at the 5 per cent level also showed significant response change at the 1 per cent level. Sixty per cent of all items showed no significant response change at the 5 per cent level in any of the four sequences and 79 per cent of all the items showed no significant response change at the 1 per cent level in any of the four sequences. Also, there was no tendency for the items which showed significant response change to be located in any particular section of the inventory.

6. The group of graduating seniors, Spring and Summer 1947, from which the To group was drawn, was divided into three major curricular groupings:
1. Early childhood education majors (nursery, kindergarten, primary, elementary).
 2. Academic field majors (English and speech, foreign language, mathematics, science, social studies).
 3. Special field majors (art, home economics, industrial, music, physical education).

After having first determined that there was no sex difference in TAI scores in the sub-groups, it was found that there were significant differences in TAI-scores among the three sub-groups with the early childhood education majors scoring highest and

the special field majors scoring lowest (see Tables 6 & 7). An inspection of TAI scores of the Jo group (first-Quarter juniors) by the same curricular sub-divisions revealed differences among the sub-groups of about the same order of magnitude as those observed for the graduating seniors (see Table 6).

TABLE 6
Mean TAI Scores for Beginning Juniors (Jo Group) and Graduating Seniors (To Group)
Classified by Major Curricula

	Jo Group		To Group	
	N	Mean	N	Mean
Sub Group I (Natural, Kindergarten, Pre- minor Elements)	49	137.19	43	165.65
Sub Group II (Academic Fields)	75	138.79	76	148.08
Sub Group III (Special Fields)	41	133.84	71	149.59
Total	175	140.11	191	148.01

TABLE 7
Analysis of Variance of TAI Scores of Three Curricular Sub Groups of Graduating Seniors, 1947

Source of Variation	df	Sum of Squares	Mean Square	F	P
Between	2	36.949	18.474	19.81	<.01
Within	191	42.171	.221		
Total	193	79.120			

TABLE 8
A Comparison of TAI Scores of the Graduating Seniors, 1947, Who Were Teaching the Following Subjects with the Scores of Those Who Were Not

Group	N	SD	Mean
Teaching	137	16.34	147.43
		$T = 1.17$ $P > .05$	$T = 9.40$ $P > .05$
Not teaching	48	20.58	145.74

Minor Findings. 1. TAI-scores and scores on a measure of intelligence (*Miller Analogies*, Form A) correlated 0.08. The sample was the group of beginning juniors.

2. The chance-half reliability coefficient for the whole Inventory was found to be 0.88 both by Product-Moment and the appropriate maximum likelihood estimate formula.

3. The test-retest reliability of the Inventory was found to be .84 by the product-moment formula and 0.80 by the appropriate maximum likelihood estimate formula.

4. A significant increase in TAI-scores ($P < .01$) was observed

in the control sequence, test-retest at a week to ten-day interval. This phenomenon is not unusual for instruments similar to the one used here.

5. There was no significant difference in TAI-scores of those graduating seniors who taught during the school year following graduation and those who did not (see Table 8).

Conclusions and Interpretation

Two major conclusions may be drawn from the findings of this investigation. The first conclusion is that *the attitudes measured by the Teacher Attitude Inventory are of sufficient stability to warrant further investigation as to their efficiency in predicting teacher-pupil relations and in pre-training selection of teacher*. Several of the findings support this conclusion. The changes in TAI scores that occurred during the time spans studied, even though significant in two of the three sequences, were not of great magnitude. Further, the increase in TAI-scores that occurred in the junior sequence was practically negated by the decrease in the teacher sequence so that by the time a teacher has taught six months his attitudes toward pupils as measured by the *Teacher Attitude Inventory* are about the same as when he began professional training as a junior. Also the correlation coefficients between first and second testing in the sequence tend to be just significantly less than the test-retest reliability coefficient, indicating that the individuals tend to hold their respective ranks during each of the time spans studied. The fact that the Inventory was found to be only slightly susceptible to attempts to fake good gives added confidence to the conclusion that the attitudes being measured are rather stable. Finally the fact that 79 per cent of the individual attitudes (items) were not affected significantly (at the 1 per cent level) by training, experience, or test-retest at a week interval, permits this conclusion to be drawn with a high degree of confidence.

The second major conclusion to be drawn from this investigation is that *there are significant differences in teacher-pupil attitudes among subjects classified by their major curriculum and that these differences are present in about the same magnitude at the beginning of professional training as at the end of it, with the*

early childhood education major ranking highest as a group and the special field majors ranking lowest as a group. This is particularly significant in view of the fact that the scoring of the items was determined on groups of teachers distributed throughout all twelve grades, so that responses of elementary grade teachers were not unduly weighted in the development of the scoring key.

It would appear that the attitudes measured by the Inventory are rather well formed by the time the subject enters pre-professional training and are influenced to only a minor extent by training and the first half year of teaching. However, there is a small group of attitudes that are affected significantly by training and another group, still smaller, that is significantly affected by experience. Also, it would appear that teacher-pupil attitudes are operative in the subject's selection of the field of education in which he wishes to specialize. It is conceivable that these attitudes have elements in common with vocational interest and that a measure of them would be useful in counseling students about vocational choices.

REFERENCES

1. Bordin, Edward S. "A Theory of Vocational Interest as Dynamic Phenomena." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, III (1943), 49-65.
2. Callis, Robert. "Change in Teacher-Pupil Attitudes Related to Training and Experience." Unpublished Ph.D. Thesis on file in the University of Minnesota Library, 1948.
3. Cook, Walter W. and Leeds, Carroll H. "Measuring the Teaching Personality." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, VII (1947), 399-410.
4. Cook, Walter W., Leeds, Carroll H. and Callis, Robert. "Predicting Teacher-Pupil Relations." *The 1949 Yearbook of the Association for Student Teaching*. Chapter IV. To be published.
5. Fischer, Robert P. and Andrews, Avonne L. "A Study of the Effect of Conformity to Social Expectancy on Evaluative Attitudes." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, VII (1947), 331-335.
6. Gough, Harrison G. "Simulated Patterns on the MMPI." *Journal of Abnormal and Social Psychology*, XLII (1947), 215-225.
7. Kelley, F. L., Miles, C. C. and Terman, L. M. "Ability to Influence One's Score on a Pencil-and-Paper Test of Personality." *Character and Personality*, IV (1936), 206-215.
8. Kimber, J. A. M. "The Insight of College Students Into the Items on a Personality Test." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, VII (1947), 411-420.

9. Leeds, Carroll H. "The Construction and Differential Value of a Scale for Determining Teacher-Pupil Attitudes." Unpublished Ph.D. Thesis on file in the University of Minnesota Library, 1946.
10. Leeds, Carroll H. and Cook, Walter W. "Construction and Differential Value of a Scale for Determining Teacher-Pupil Attitudes." *Journal of Experimental Education*, XVI (1947), 149-159.
11. Meehl, Paul E. and Hathaway, S. R. "The K-Factor as a Suppressor Variable in the MMPI." *Journal of Applied Psychology*, XXX (1946), 525-564.
12. Ruch, Floyd L. "A Technique for Detecting Attempts to Fake Performance on the Self-Inventory Types of Personality Test." *Studies in Personality* Contributed in Honor of Lewis M. Terman. New York: McGraw-Hill Book Company, 1942.
13. Steinmetz, H. C. "Measuring Ability to Fake Occupational Interest." *Journal of Applied Psychology*, XVI (1932), 123-130.
14. Strong, E. K. *Vocational Interests of Men and Women*. Stanford: Stanford University Press, 1943.

A STUDY OF CLIENT RESPONSIBILITY: COUNSELOR TECHNIQUE OR INTERVIEW OUTCOME?

CHARLES F. HILTON

Ohio State University

The recent advances in interview methodology illustrate the fruitfulness of an empirical approach to the complex relationships within the interview (1, 6, 7, 8, 10). However, it is obvious that many pertinent problems remain to be investigated. One important area of interest is the relationship between counselor technique and interview outcome. Related to this area of interest is the current controversy between "directive" and "non-directive" therapy and the questions which have been raised concerning the role of client responsibility during the interview.

An important goal or outcome of counseling is the assumption by the client of self-responsibility (2, 3). Consequently, the degree to which a client takes self-responsibility has been used as a criterion of interview effectiveness. However, forcing a client to take responsibility for the direction of the interview is a commonly used counselor technique (5, 9). The purpose of this study is to clarify this dual role of responsibility-taking during the interview. We shall seek to answer the following questions. (1) Is the assumption of responsibility by a client during an interview an important outcome or criterion of interview effectiveness? (2) Does the throwing of responsibility upon a client represent a useful counseling technique? And (3) may the responsibility behavior of a client during an actual interview be differentiated into that responsibility which is a criterion and that responsibility which is a manifestation of counseling technique?

Description of the Data

The data used in this study were obtained from transcripts of 78 interviews which were recorded in a counseling practicum offered for advanced students at The Ohio State University.

Of these interviews 42 were conducted by 7 experienced counselors while the remainder were held by 16 counselors-in-training. For a period of ten weeks these counselors met weekly with students in a how-to-study course. Although the problems of the counselees generally revolved around study difficulties, other problems of individual adjustment were frequently encountered. Characteristically, the students in this course are normal individuals who desire to become more effective in their college relationships (4).

Definition of Variables

The unit of analysis used in this study consists of the counselor and counslee remarks which are related to a particular problem as it is discussed in the interview and is called the "discussion unit." Previous interview analysis has shown that there is a high degree of reliability between judges in differentiating the end of one topic of conversation and the beginning of a new topic (7). This sort of unit analysis produced 421 discussion units in the 78 interviews. However, 68 of these were not used because they were too short for reliable classification or dealt with such special topics as social visiting, making suitable arrangements, etc. The units used were concerned with four general topics, namely, study skills, scholastic questions, vocational problems and personality problems. Because the last three of these four topics all dealt with making decisions in different fields, an initial analysis was made to determine if these topics showed similar distributions. The results indicated the feasibility of combining the topics of vocational problems, scholastic questions, and personality problems under a general classification of decision-making units. Thus, in the 353 interview units there were 148 study-skill and 205 decision-making units.

The concept of responsibility-taking refers to the degree to which the client or counselor is responsible for the direction of the interview. Within an interview the division of responsibility may rest, at one extreme, with the counselor, at the other extreme with the client, or the responsibility may be shared to a large degree between the counselor and client. This division of responsibility between counselor and client during each dis-

discussion unit was rated on a five-point scale—1 representing the highest degree of counselor responsibility and 5 representing the highest degree of counselee responsibility.

Since a counselor usually determines responsibility-assigning through the type of remarks he makes, it was also important to study these types of counselor remarks. The most important characteristic of counselor remarks which determines responsibility is their degree of "leading." By leading is meant the degree to which each counselor's remark seems to go beyond the thinking of the client as expressed in his preceding statement. The concept of leading does not necessarily mean pulling the client along; it may represent a team-like arrangement—as between a pass thrower and receiver in football—in which the counselor speaks in terms of what he thinks the client is ready for.

Thus, the two important characteristics of a particular counselor technique are the amount of lead and the degree to which the counselor is responsible for the direction of the interview. For example, if a counselor clarifies a client's previous remark for the client, an often used technique, the counselor is not introducing any new idea into the interview and, hence, in terms of relative distance, he has not moved ahead of the client's thinking but at the same time he has thrown the responsibility for the next point to be discussed onto the client. (It is to be noted, however, that in using clarification the counselor is leading in the sense that he may select one idea out of many to respond to.) Suppose, however, that a counselor urges a client to undertake a certain course of action. In this instance the counselor may have introduced a new idea to the client and if the client does not possess enough insight to understand the relation of this idea to his problem and is not ready for the idea, the counselor is quite some distance ahead of the client's thinking. Here the counselor is assuming, at the same time, responsibility for the point brought up in the interview.

The "primary counselor technique" for a discussion unit was determined by tabulating each counselor response within a discussion unit according to 1 of 10 categories of leading techniques. Within any interview unit the modal technique was called the primary counselor technique. The result of this tabu-

lation indicated that clarification, tentative analysis, interpretation, and urging were the most frequently used counselor techniques. Only those units employing one of these techniques were used in the present study. Detailed definitions of these techniques are given in a recent article by Carnes and Robinson (1).

Evidence of interview outcomes consisted of ratings of (a) growth in counsellee insight during the discussion unit, (b) the working relationship between counselor and client, and (c) the division of responsibility for the direction of the interview. Five-point rating scales were used for each of these variables. Sherman found that their reliability was sufficiently high to justify their use with this type of problem (7).

Results

1. What relationship does division of responsibility have to interview outcome? One means of answering this question is to correlate responsibility ratings with other ratings of known interview outcomes, e.g., insight and working relationship. Product moment correlation coefficients were obtained separately for 148 study skill units and 205 decision making units. These correlations were of the order of .49 between responsibility and insight for the study skill units and .51 between responsibility and insight for the decision making units; .47 between responsibility and working relationship for the study skill units and .37 between responsibility and working relationship with the decision making units. The possibility of chance factors contributing to these correlations may be rejected at the one per cent level of confidence. While a correlational relationship has been shown, no assumption may be made from these coefficients as to the linearity of the relationship nor to the factor or factors involved.

Of further interest is a qualitative analysis of the average insight and working relationship ratings for each responsibility rating (Table 1). It will be recalled that a responsibility rating of 1 signifies the highest degree of counselor acceptance of responsibility for the direction of the interview, while a rating of 5 signifies the highest degree of client acceptance of responsibility for the direction of the interview. In the case of insight

and working relationship a rating of 1 means no insight or a resistive client, while a rating of 5 means a high degree of client self insight or an excellent working relationship between client and counselor.

The following generalizations are drawn from Table 1: (a) Counselor insight is not likely to develop when the counselor assumes, through choice or necessity, the entire responsibility for the direction of the interview. (b) The middle ratings of responsibility tend to be accompanied by the greatest gains in stated insights. (c) A harmonious relationship between the client and counselor is most likely to be obtained if the responsibility for the direction of the interview is shared between the counselor and client. The averages reported in Table 1 are not

TABLE 1
The Mean Insight and Working Relationship Ratings for Units with Each Level of Responsibility

	Responsibility				
	1	2	3	4	5
Study Skill Units (N = 148)					
Insight	1.52	2.31	3.04	3.66	4.09
Working Relationship	3.11	3.84	4.47	4.11	3.65
Decision Making Units (N = 200)					
Insight	1.33	2.19	2.55	3.93	4.14
Working Relationship	2.87	3.41	4.13	4.17	4.11

amenable to refined interpretation, but it needs to be pointed out that the relationship between responsibility and the other outcomes of insight and working relationship is not strictly a linear one in the sense that high values in one outcome are accompanied by correspondingly high values in the other. In other words, the middle range of these rated outcomes is generally most effective in promoting positive relationships within the interview.

2. Is the throwing of responsibility upon a client a useful counseling technique? Having found that responsibility-taking is positively related to insight and working relationship, we may now investigate the influence of counselor technique upon the degree of responsibility assumed by the client. In order to do this the 353 interview units were divided into groups: first, on the basis of topic of conversation, i.e., either study skill or decision-making, and second, on the basis of primary counselor

technique used, i.e., clarification, tentative analysis, interpretation, and urging. With the data thus classified the mean responsibility was computed for all of the units characterized by the use of a given primary counselor technique. These mean values are shown in Table 2. Subsequently, the significance of the differences between the mean responsibility values was determined by an analysis of variance. The results are reported in the following two paragraphs.

When study skills was the topic of conversation the variance within the primary counselor techniques was .116, which is significant at the 5 per cent level of confidence. However, between the individual techniques some further differences were found. The variance difference between the means of clarification and tentative analysis could have occurred by chance alone. This was also true for the difference between clarification and

TABLE 2
*The Mean Responsibility Ratings for 128 Units of Conversation Making Units
Divided According to the Primary Counselor Technique Used*

	Clarification	Tentative Analysis	Interpretation	Urging
Study Skill Units	2.42	2.59	2.14	0
Decision Making Units	2.60	2.22	2.14	0.17

interpretation. The difference between clarification and urging was 10.8%, between tentative analysis and urging it was 15.6%, both of which are significant at beyond the one per cent level of confidence. The variance between the means of tentative analysis and interpretation was .476, between interpretation and urging it was 6.51, both of which are significant at the 5 per cent level of confidence. The foregoing values indicate that the degree of responsibility assumed by the counsellee is affected by counselor technique and, particularly, that the use of urging as a counselor technique is likely to keep the counsellee from taking responsibility.

When decision making units were the topic of conversation the variance within all of the primary counselor techniques was .1719, which is significant at beyond the one per cent level of confidence. The variance between the means of clarification and tentative analysis could have occurred by chance alone and, hence, is not significant. The variance between the means

of clarification and interpretation was 15.05, between the means of clarification and urging it was 39.05, between the means of tentative analysis and interpretation it was 12.39, between tentative analysis and urging it was 35.95, and between interpretation and urging it was 14.80. All of these values are significant at beyond the one per cent level of confidence. Hence, again, the evidence points to the fact that the primary counselor technique affects the degree of responsibility assumed by the client and that every technique is superior to urging in obtaining client responsibility-taking.

Although the analysis above provides, in some instances, clear-cut distinctions between the counselor techniques it does not warrant too broad generalizations about the effectiveness of these primary counselor techniques. That is, mean values are given in Table 2; each primary technique had units with ratings of the highest effectiveness in both working relationship and insight. Each technique may be useful to the well-trained counselor, primarily their usefulness is a function of the situation in which the counselor is forced to operate. It is entirely conceivable that with certain types of clients and problems a counselor could use urging with effective results.

3. Is it possible to differentiate the responsibility behavior of a client during an actual interview into that which is client initiated, i.e., a criterion, and that which is simply a manifestation of counseling technique? In the first part of this paper it was shown that responsibility-taking was correlated with insight and working relationship. However, our analysis of variance has shown that technique of leading is highly related to the responsibility assumed by the client. While the first result might be used as an argument that responsibility taking is an outcome, the second result shows that responsibility-taking is markedly affected by counselor technique. It is necessary to know whether responsibility-taking is an outcome when the effect of counselor technique is controlled. This control was achieved in two ways.

First, responsibility was correlated with insight and working relationship separately for each set of units having a common counselor technique and the same subject matter. These correlations are shown in Table 3. It will be seen that when the

influence of counselor technique is held constant in this manner responsibility-taking is still positively correlated in each case with insight and working-relationship.

Second, some general means are needed to measure responsibility-taking as an outcome without the necessity of presenting it in so many sub-divisions. This may be done through the use of "derived scores." These derived scores were computed by subtracting the mean responsibility rating for all of the units of a given topic and counselor technique from each obtained rating of responsibility in such units. As an example, suppose that our mean responsibility value for decision making units in which interpretation is the primary counselor technique

TABLE 1
Correlation of Responsibility with Working Relationship and with Insight for Units Characterized by Four Primary Counselor Techniques

	Classification	Interpretation Analysis	Interpretation Unit	Rating	Derived Score
Study-Skill Units					
a	Responsibility correlated with insight	26	42	37	46
b	Responsibility correlated with working relationship	26	40	38	46
Decision Making Units					
a	Responsibility correlated with insight	10	31	24	42
b	Responsibility correlated with working relationship	10	29	26	42

was 2.40. In computing the derived score, any such unit classified with a responsibility rating of 3 obtained a derived score of plus .60, with a rating of 2, a derived score of minus .40, etc. These derived scores function to control the variable influence of counselor technique which is how to influence responsibility-taking. That is, it is assumed that clients with positive derived scores wanted to take more responsibility than was allowed or forced upon them by the counselor technique, while those with negative scores did not want to assume responsibility.

Using derived scores the following coefficients of correlation were obtained: between responsibility and working relationship with study-skill units, .46, between responsibility and insight with study-skill units, .37, between responsibility and working relationship with decision making units, .52, between responsibility and insight with decision-making units, .62. These cor-

relations represent the relationship of responsibility-taking to other interview outcomes when the effect of counselor technique has been minimized, and they substantiate our former conclusion that responsibility-taking is an important interview outcome or criterion.

It is believed that the concept of derived score is an important contribution to future research. If, as has been suggested elsewhere, responsibility-taking is a major dimension of the interview process, a means is now available to further clarify the complexity of this dimension. Certainly, in planning counselor training programs, knowledge is needed about the relative effectiveness of counselor techniques in producing desirable interview outcomes. The derived score will function in such research to control the effect of counselor technique while the variable relationships among the interview outcomes are investigated.

Two cautions need to be exercised in the interpretation of the analysis in this study. Responsibility, insight, and working relationship are symptoms of goals in counseling, e.g., increased happiness, effectiveness, etc., and are not simply to be sought in themselves. Because delayed goals cannot be measured in the moment-to-moment work of an interview these immediate criteria are useful tools in judging counseling progress during an interview. Further, responsibility-taking might be an independent outcome of counseling, our argument has been proof by similarity only.

Summary and Conclusions

An important goal of counseling is to aid the client to become responsible for his behavior. However, forcing a client to take responsibility is a commonly used counselor technique. This dual role of responsibility-taking was investigated. Specifically, an analysis was made of the relationship between the degree of client responsibility within the interview and the variables of (1) interview outcome, e.g., growth in counselee insight and working relationship and, (2) counselor technique.

The data used were taken from 78 recorded interviews which occurred in conjunction with a how-to-study course offered at The Ohio State University. The 78 interviews were broken down into 353 discussion-topic units.

Responsibility as an interview outcome, and as a result of counselor technique, was analyzed in three ways: (1) Ratings of responsibility were correlated with other interview outcomes, (2) The effect of technique upon responsibility was determined by an analysis of variance of the mean differences of each primary counselor technique, (3) A differentiation was made between the effect of technique and outcome by the use of derived scores.

It is concluded that: (1) The amount of responsibility assumed by a counsellee can be affected by the counselor technique used and (2) when the effect of counselor technique is controlled, responsibility-taking is related to other interview outcomes and, as such, it may be used as a criterion of interview effectiveness.

REFERENCES

1. Carnes, E. F. and Robinson, F. P. "The Role of Client Talk in the Counseling Interview." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, VIII(1948), 635-644.
2. Darley, J. G. *The Interview in Counseling*. Washington, Dept of Labor, Retraining and Re-employment Administration, 1946.
3. Pennington, L. A. and Berg, I. A. *An Introduction to Clinical Psychology*. New York: Ronald Press, 1948. Chapter 18.
4. Robinson, F. P. "Two Queries with a Single Stone." *Journal of Higher Education*, XVI (1945), 201-206.
5. Rogers, C. R. "Psychometric Tests and Client-Centered Counseling." *EDUCATION AND PSYCHOLOGICAL MEASUREMENT*, VI(1946), 139-144.
6. Seeman, J. "A Study of Client Self-Selection of Tests in Vocational Counseling." *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, VIII(1948), 327-346.
7. Sherman, D. M. "An Analysis of the Dynamic Relationships Between Counselor Technique and Outcome in Larger Units of the Interview Situation." Unpublished Doctor's Dissertation, Ohio State University, 1943.
8. Snyder, W. U. "An Investigation of the Nature of Non-Directive Psychotherapy." *Journal of General Psychology*, IV(1948), 256-263.
9. Thorne, F. C. "Further Critique of Non-Directive Methods of Psychotherapy." *Journal of Clinical Psychology*, XXXIII (1945), 193-223.
10. Tindall, R. H. and Robinson, F. P. "The Use of Silence as a Technique in Counseling." *Journal of Clinical Psychology*, III(1947), 136-141.

RECENT PUBLICATIONS RECEIVED

- ABT, LAWRENCE I. and BELLAK, LEOPOLD (EDITORS) *Projective Psychology* York: Alfred A. Knopf, 1950. 485 pp. \$6.00.
- BERGER, GASTON *Traité Pratique d'Analyse du Caractère* Paris: Presses Universitaires de France, 1950. 250 pp. 500 fr.
- Buros, OSCAR K. (CHAIRMAN) *Institutional Conference on Testing Problems* Princeton: Educational Testing Service, 1950. 94 pp.
- CATTELL, RAYMOND B. *Personality: A Systematic Theoretical and Factual Study*. New York: McGraw-Hill Book Company, 1950. 689 pp. \$5.50.
- DRUCKER, ARTHUR J. *Further Studies in Attitudes, Series XVI, Relationships Between Citizenship Attitudes, Parental Education, and Other Variables* Lafayette: The Division of Educational Reference, Purdue University, 1950. Studies in Higher Education LXXI, 63 pp. \$75.
- EDUCATIONAL RECORDS BULLETIN No. 54. *1950 Achievement Testing Program in Independent Schools and Supplementary Studies*. New York: Educational Records Bureau, 1950. 119 pp.
- FARNSWORTH, PAUL RANDOLPH *Musical Taste: Its Measurement and Cultural Nature*. Stanford: Stanford University Press, 1950. 94 pp. \$1.50.
- GUILFORD, J. P., COMBEE, A. J., R. F. and CHRISTENSEN, P. R. *A Factor-Analytic Study of Reasoning Abilities I. Hypotheses and Description of Tests*. Beverly Hills: The University of Southern California. Reports from the Psychological Laboratory, No. 1, 1950. 23 pp.
- HAHN, MILTON F. and MACLEAN, MALCOLM S. *General Clinical Counseling in Educational Institutions*. New York: McGraw-Hill Book Company, 1950. 375 pp. \$3.50.
- HARDEN, EDGAR L. *How to Organize Your Guidance Program*. Chicago: Science Research Associates, 1950. 70 pp.
- HOBSON, ROBERT L. *Further Studies in Attitudes, Series XVIII, Some Psychological Dimensions of Academic Administrators* Lafayette: The Division of Education Reference, Purdue University, 1950. Studies in Higher Education LXXIII, 99 pp. \$1.50.
- LANDIS, CARNEY, and BOLLES, M. MARJORIE. *Textbook of Abnormal Psychology, Revised Edition*. New York: The Macmillan Company, 1950. 634 pp. \$5.00.
- LE GALL, ANDRÉ. *Caractérologie des Enfants et des Adolescents*. Paris: Presses Universitaires de France, 1950. 458 pp. 800 fr.
- PORTEUS, STANLEY D. *The Porteus Maze Test and Intelligence*. Palo Alto: Pacific Books, 1950. 194 pp. \$4.00.

- REMMERS, H. H. (EDITOR) *Further Studies in Attitudes, Series XV, Studies in College and University Staff Evaluation*, Lafayette. The Division of Educational Reference, Purdue University, 1950. Studies in Higher Education LXX. 99 pp. \$1.25.
- RUPE, JESSE C. *Further Studies in Attitudes, Series XVIII, Some Psychological Dimensions of Business and Industrial Executives* (Bound with Hobson, above).
- SEASHORE, ROBERT H. and VAN DUSEN, A. C. *How to Solve Your Problems* (Life Adjustment Series). Chicago: Science Research Associates, 1950. 48 pp. \$.60.
- SHIMBERG, BENJAMIN. *Further Studies in Attitudes, Series XVII, The Development of a Needs and Problems Inventory for High School Youth*. Lafayette: The Division of Educational Reference, Purdue University. Studies in Higher Education LXXII. 78 pp. \$1.25.
- WRENN, C. GILBERT and DUGAN, WILLIS E. *Guidance Procedures in High School*. Minneapolis: The University of Minnesota Press, 1950. The Modern School Practices Series, Number One. 71 pp.

THE CONTRIBUTORS

B. J. Borreson B A , University of Minnesota, 1945. Director, Student Housing Bureau, 1946-1948, Associate Director, Student Activities Bureau, 1948- , University of Minnesota. Member of the Subcommittee of the Student Personnel Work Committee, charged with preparing a brochure on education and student life. Panel member, Minnesota Governor's State Conference on Youth.

Robert Callis Ph D , University of Minnesota, 1948. With the U S Navy, 1942-1946. Counselor, General College, University of Minnesota, 1946-1948. Assistant Professor of Psychology and Head of Counseling Bureau, 1948-1950, Associate Professor of Psychology and Head of Counseling Bureau, 1950 , University of Missouri. Member, American Psychological Association, American College Personnel Association, Psi Chi, Phi Delta Kappa.

Louis D. Cohen Ph D , Duke University, 1949. Head Psychologist, New York City Penitentiary, 1937-1938. Director of Classification and Education, Indiana State Farm, 1938-1942. Lieutenant-Lieutenant-Colonel, U S Army, 1942-1946. Associate in Clinical Psychology, 1946-1949, Assistant Professor of Psychology and Assistant Professor of Neuropsychiatry, 1949 , Duke University. Fellow, American Psychological Association, North Carolina Psychological Association. Member, Sigma Xi, Southern Society of Philosophy and Psychology, North Carolina Academy of Science, SPSSI. Diplomate in Clinical Psychology, American Board of Examiners in Professional Psychology.

N. M. Downie -Ph D, University of Syracuse, 1948. Instructor in Biology, Robert College, Istanbul, Turkey, 1936-1939. Instructor in Education and Graduate Assistant, Evaluation Service Center, Syracuse University, 1946-1948. Assistant Professor of Education, State College of Washington, 1948-

Paul L. Dressel—Ph D , University of Michigan, 1939. Director of Counseling and Chairman, Board of Examiners, Michigan State College, 1944-. Author of articles in the field of measurement. Member, American Educational Research Association, American Psychological Association, American College Personnel Association, Institute of Mathematical Statistics, National Vocational Guidance Association, Psychometric Society, American Association for the Advancement of Science, American Association of University Professors, Sigma Xi, Phi Delta Kappa.

Albert Ellis—Ph D , Columbia University, 1947. Consulting Psychologist, New York City, 1943-. Senior Clinical Psychologist,

Mental Hygiene Clinic, New Jersey State Hospital at Greystone Park, 1947-1948. Lecturer in Psychology, Rutgers University, 1948-1949. Lecturer in Mental Hygiene, New York University, 1949-. Chief Clinical Psychologist, The Diagnostic Center, Menlo Park, N. J., 1949-. Author of articles in educational, psychiatric, and sociological journals. Member, American Psychological Association, Eastern Psychological Association, American Orthopsychiatric Association, American Sociological Society, American Association of Marriage Counselors, New York Association of Clinical Psychologists, New Jersey State Psychological Association, New York Academy of Science, American Association on Mental Deficiency, American Association for the Advancement of Science, National Committee for Mental Hygiene, National Mental Health Foundation, Kappa Delta Pi, Phi Delta Kappa.

Charles F. Elton--M. A., Ohio State University, 1948. Assistant Instructor, Department of Psychology, Ohio State University, 1950-

Martin F. Fritz--Ph. D., University of Chicago, 1931. Instructor, Assistant Professor, Associate Professor, Professor of Psychology, 1927-. Clinician, Testing Bureau, 1944-. Iowa State College. Author of articles on learning, psychodietetics, and tests. Member, American Psychological Association, Midwestern Psychological Association, Iowa Psychological Association, Iowa Academy of Science, Iowa Society for Mental Hygiene, Phi Kappa Phi, Pi Kappa Delta, Phi Delta Kappa, Psi Chi.

Edward J. Furst--Ph. D., University of Chicago, 1948. Classification Specialist and later, Classification Officer, Adjutant General's Department, U. S. Army, 1942-1946. Research Assistant, Board of Examinations, University of Chicago, 1947. Assistant Chief, Evaluation and Examinations Division, Bureau of Psychological Services, University of Michigan, 1948-. Associate Member, American Psychological Association. Member, American Educational Research Association, Phi Delta Kappa, Phi Beta Kappa.

Robert R. Holt--Ph. D., Harvard University, 1944. Research Assistant, Harvard Psychological Clinic, 1941-1944. Study Director, Division of Program Surveys, B. A. E., U. S. Department of Agriculture, 1944-1946. Clinical Psychologist, Winter V. A. Hospital, Topeka, Kansas, 1946-1949. Associate Psychologist, Senior Psychologist, The Menninger Foundation, 1947-. Instructor, Menninger School of Psychiatry, 1946-. Editor, TAT Newsletter, 1946. Author of articles in professional journals. Associate member, Topeka Psychoanalytic Society. Fellow, American Psychological Association (Chairman, Committee on Publication Outlets in Clinical Psychology). Member, Group of Psychoanalytic Psychologists, Society for Projective Techniques (Editorial Advisory Board), Kansas Academy of Sciences, American Association for the Advancement of Science, Sigma Xi, Phi Beta Kappa.

Robert W. Irvine B.A., Cornell College, I.I.B., University of Minnesota, 1948 Counselor, 1946-1948, Associate Director, Student Activities Bureau, 1948-1949, University of Minnesota Practising Attorney, Detroit Lakes, Minnesota, 1949 Member, Hennepin County Bar Association, Seventh Judicial District Bar, Minnesota Bar Association

A. R. Lauer Ph.D., Ohio State University, 1929. Member of the Department of Psychology, Iowa State College, 1925- Author of *Learning to Drive Safely*, co-author of *The Driver—His Nature and Improvement* and author of articles on research of a scientific nature. Member and past member of President's Highway Safety Conference, National Research Council, Highway Research Board, National Safety Council, American Optometric Association, National Education Association, Governor's Highway Safety Commission in Iowa, Iowa Safety Council, Iowa Safety Congress Consultant on driving and drivers to Appalachian Electric Power and Light Co., The Philadelphia Co., Iowa Highway Patrol, Borden Milk Co., Ruan Transfer Co., United Motor Coach Co., Bell Telephone Co., Ford Motor Co., Chrysler Motor Co., and other transportation companies and organizations. Member, American Psychological Association, Midwestern Psychological Association, Sigma Xi, Phi Delta Kappa, Phi Chi and numerous other professional and honorary groups

Ross W. Matteson—Ph.D., University of Michigan, 1949 Director, Guidance and Research, Hazel Park High School, 1939-1943 With the U.S. Navy, 1943-1946 Counselor, Michigan State College Counseling Center, 1946- Member, American College Personnel Association, Phi Delta Kappa, National Vocational Guidance Association, American Psychological Association.

William B. Michael—Ph.D., University of Southern California, 1947 Teaching Assistant in Mathematics, 1942-1943; Instructor, Engineering Mathematics, E.S.M.W.T., 1942-1945, California Institute of Technology Instructor in Mathematics, Pasadena Junior College, 1943-1944 Lecturer in Mathematics, 1944-1945; Lecturer in Education and Psychology, 1945-1947, University of Southern California. Assistant Professor of Psychology, Princeton University, 1947- Associate Member, American Psychological Association Member, Mathematical Association of America, Institute of Mathematical Statistics, American Statistical Association, Psychometric Society, Western Psychological Association, Southern California Psychological Association, Phi Beta Kappa, Sigma Xi, Phi Kappa Phi, Phi Delta Kappa

R. C. Myers—Ph.D., Stanford University, 1950. Social Science Analyst, U. S. Army, 1942-1945. Faculty member: Stanford University, 1945; University of Oregon, 1945-1946, Princeton University, 1947-1949. Research Associate, Audience Research, Inc., 1946-1947. Research Associate to Project Director, Educational Testing Service,

1947- Author of articles on social psychology, attitude and opinion research methodology Member, American Sociological Society, Eastern Sociological Society, SPSSI

Charles O. Neldt—Ph D, Iowa State College, 1949 Instructor, Assistant Professor, Iowa State College, 1948- Author of articles on tests and test instruction Associate member, American Psychological Association Member, Midwestern Psychological Association, Iowa Psychological Association, Iowa Academy of Science, Phi Kappa Phi, Phi Delta Kappa, Psi Chi.

C. Robert Pace—Ph D, University of Minnesota, 1937 Instructor and Research Associate, General College, University of Minnesota, 1937-1940 Research Associate, Commission on Teacher Education, American Council on Education, 1940-1943 Head, Research Unit and Field Research Section, Bureau of Naval Personnel, 1943-1947 Associate Director, Director, Evaluation Service Center, Syracuse University, 1947- Author of *They Went to College* (Univ of Minnesota Press) and co-author of *Evaluation in Teacher Education* (American Council on Education); author of articles on attitude measurement, evaluation, and higher education Fellow, American Psychological Association Member, American Educational Research Association, National Society for the Study of Education, American Association for Public Opinion Research

D. G. Schultz—Ph D, Pennsylvania State College, 1945 Research Assistant, Personnel Service Division, 1943-1944; Instructor in Psychology, 1944-1945, Pennsylvania State College Assistant Head, Test Construction Department, 1946-1948, Research Associate, 1948-, Educational Testing Service Associate Member, American Psychological Association Member, Eastern Psychological Association, Psychometric Society, Sigma Xi

Maurice E. Troyer—Ph D, Ohio State University, 1935. Superintendent, Bureau of Township Schools, Princeton, Illinois, 1925-1929 Assistant Professor of Psychology, Bluffton College, 1930-1932 Instructor in charge of Remedial Program, Ohio State University, 1933-1936 Assistant Professor of Education, Syracuse University, 1936-1939 Associate Professor, 1939 Associate in Evaluation, Commission on Teacher Education, American Council on Education, 1940-1943 Director, Bureau of School Services, Professor of Education, Syracuse University, 1943 Director, Evaluation Service Center, Syracuse University, 1945 Member, American Psychological Association, American Association of Applied Psychology, American Educational Research Association, American Association for the Advancement of Science.

